Character and Culture in Social Cognition

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

2022

James C. Lloyd

The School of Social Sciences

The Department of Philosophy

Blank page

Table of Contents

Introduction: Reading Mind and Character

Chapter One: Cognising Social Cognition

<u>Final word count:</u> 70809

Abstract:

We make character trait attributions to predict and explain others' behaviour. How should we understand character trait attribution in context across the domains of philosophy, folk psychology, developmental psychology, and evolutionary psychology? For example, how does trait attribution relate to our ability to attribute mental states to others, to 'mindread'? This thesis uses philosophical methods and empirical data to argue for character trait attribution as a practice dependent upon our ability to mindread, which develops as a product of natural selection acting on culture instead of genes. This analysis carves out trait attribution's distinct place within an emerging complex and mature scholarship on pluralistic social cognition.

Declaration:

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement:

Dedications:

This dissertation has been my life for the last three and a half years—it absolutely would not have been completed on time and to the quality that I envisaged without the input and support from a great many people. This is particularly in light of the Coronavirus-19 pandemic.

The most important person to thank is my supervisor, Joel Smith, for his unwavering and dedicated support. Without his numerous, constant, insightful, and incisive comments, without the hours spent in conversation, without his guiding hand of expertise and philosophical acumen, this thesis would not have been anywhere near as philosophically rigorous as it is today. I am also very grateful to my secondary, Michael Scott, for helpful advice during reviews. An important thanks is also due to the funding committee at The University of Manchester, who gave me the opportunity to spend these years thinking and writing.

I would also like to thank my friends at The University of Sheffield for their support and for many valuable philosophical conversations over the years. In particular, I would like to thank Stephen Laurence, for his care and mentorship during my undergraduate and master's degrees—I am here because of his direct encouragement. Thanks also to Rosa Vince for support in the early stages of undertaking the thesis, and also to everyone who read drafts of my work—especially my long-suffering proof-reader, Rose Cocker. Finally, I wish to thank the community of researchers in Manchester for contributing to making the last few years the most interesting of my life.

Reading Mind and Character

1. Introduction

Whilst we attribute thoughts and feelings to others to understand their actions, we also attribute character traits to them. Thinking the old man a grumpy person, I might predict a rebuke when accidentally kicking footballs into his garden. We might ask 'why is Breonna in the office so late?' where often, one's judgment that she is a hard-working person is a perfectly reasonable explanation. As such, there is a question of how one attributes the trait to the person to begin with. I do not mean that we might wonder what it is about, for example, donating to charity that makes one generous, though this is a good question.[1] I am concerned with what allows one to link/relate a person to their supposed character traits such that one can use this attribution in practical reasoning. Then, once we do this, the question is how *do* we reason with traits, i.e., how do we reach trait-based predictions or explanations of behaviour?

1.1 The Thesis Question

This dissertation is situated within the philosophy of social cognition, and more broadly within the philosophy of psychology. This dissertation is directed towards answering the following:

- How should we understand the ontology of character traits?
- How does the ability to attribute traits function when understanding others through their personality?
- How does this understanding relate to our abilities to understand the thoughts and feelings of others?

Some clarifications: 'ontology' here refers not only to the metaphysics of traits but also to the cognitive architecture and processing of how we understand others through character.[2] 'Function', the processing enacted to invoke our capacity for understanding

[1] On generosity specifically, see Miller (2018). Miller takes generosity to be of two senses—f irstly, where a large tip is generous when it is plentiful, relative to norms of tipping, and secondly where it has (primarily) been performed with an altruistic motive, where what is given is of value to the giver, where such an action is not required, *ceteris paribus* (pp. 219-227). What it is about giving to charity that makes one generous, for Miller, is when instances of giving to charity that meet the generous criterion form a stable disposition towards such behaviour.

[2] 'Cognitive architecture' is a broad term in this dissertation, used for any accounts of the structure of a particular cognitive skill, such as in its physical mechanism and the structure of its processing, or accounts of how the mind and thinking is structured generally—the term applies to various accounts which are couched at levels of description from neurons up to conscious reasoning.

others through character, is characterised here broadly in relation to another key feature of social cognition, of attributing thoughts and feelings to others: an ability known as 'mindreading'. Finally, understanding others through character encompasses both the attributions of character that we make to ourselves and others, and the reasoning processes that produce relevant socially cognitive results like predicting or explaining others' behaviour. As such, the discussion of trait attribution and reasoning together will be known as 'character reading*'.*

1.2 Why this? Why now?

Character traits have been historically discussed across the domains of virtue ethics and moral psychology. Broadly speaking, in virtue ethics they are studied in the vein of what makes a good character (Aristotle 2019), whereas in moral psychology traits may be invoked in understanding the psychology behind moral judgments (see Nadelhoffer, Nahmias and Nichols 2010, pt. 3). The focus here on social cognition—how character trait attribution functions in helping us understand others—is neither of these things, though moral cognition may develop from this generality (Lucca, Hamlin and Sommerville, 2019); my topic is thus relatively under-examined by philosophers. Recent scholarship in cognitive science (inclusive of philosophy) makes it prudent to examine the thesis questions now, particularly in light of recent literature that seeks to expand our understanding of mental-state attribution and how our other socially cognitive tools for understanding (such as trait attribution) relate to such skills.

The claims that I build towards are the following, though one should note that these might appear opaque for now—explication follows in the literature review and main chapters:

- Character reading is a socially cognitive skill that differs from 'mindreading' on theoretical, metaphysical, and empirical grounds.
- The cognitive processing involved (both conscious and unconscious) in how we character read can be captured by 'hybrid theory/simulation' accounts (which I will detail).
- Due to the situating of character reading within an emerging 'pluralistic' understanding of socially cognitive skills, of seeing how character reading happens unconsciously and consciously, and of understanding how trait attribution depends on a prior-emerging 'mindreading' skill, trait attribution should be considered a

'cognitive gadget', a cognitive mechanism that is the product of cultural, rather than genetic, evolution.

Sections two and three give a history of the literature that contextualises the writing of this dissertation. Section two helps the reader navigate the philosophical and psychological history of mindreading. Section three discusses some philosophical motivations for researching character traits, before making my stance on the existing empirical work clear. Section four guides the narrative towards situating my theoretical commitments in an empirically supported and *pluralistic* social cognition of character. Finally, Section five summarises the main goals of each dissertation chapter. Note that some definitions and contextualisations require further elaboration in upcoming chapters.

2. Contextualising Mindreading Research

The term 'folk psychology' recurs often in the study of social cognition. It receives more detailed treatment in chapter one, but for now we can understand it as the set of abilities affording humans the capability of reasoning about the behaviour and minds of others. One of our folk-psychological skills is 'mindreading'.[3] To mindread is to attribute mental states to others. For example, we might say that Anita drank the water because she desired thirst quenching and believed that drinking water would achieve this. This example demonstrates a common theme in this research, of a focus on the attribution of beliefs and desires to predict or explain behaviour. Prediction/explanation and talk of beliefs/desires are not exhaustive of mindreading. For example, we might mindread to manipulate instead of to predict whilst attributing mental states that are not propositional attitudes, such as the emotion of happiness. Nonetheless, propositional attitude attribution of the belief/desire kind has received the most attention in philosophical and psychological literature. The following is a brief history of key issues in research on mindreading, before moving to the discussion of research on character traits in the empirical sciences. This will provide context for the claims I make in this dissertation and will serve to carve out my research space.

[3] Skills are, roughly, abilities which can be inculcated (by environment or learning), and which can be improved upon. Nothing substantial turns on my claim that mindreading or character reading are skills. Neither does anything turn on a distinction between *a* skill and acting *with skill*, as both require a particular kind of knowledge called 'know-how' (Stanley and Williamson 2017); this knowledge is not taken to be a propositional form of knowledge and hence not necessarily language-involving. Stanley and Williamson's view of skills as a form of disposition is compatible with my claims about traits in Chapter two, given—as I claim—that whilst traits are not mere dispositions, they include them.

2.1. False Belief Understanding

Research on mindreading, or originally 'theory of mind', became popular after work by Premack and Woodruff (1978), who investigated whether chimpanzees could attribute the mental state of belief to others in order to make predictions of behaviour. Despite its positive findings, their work came under much philosophical scrutiny, particularly from Dennett (1978), Bennett (1978), Harman (1978), and also Pylyshyn (1978). They argued that such tests had not strictly proved the attribution of the belief concept, given that the tests could be (theoretically) passed by reasoning about mere behaviour. In order to test whether agents could attribute beliefs, it was reasoned, one needs to reason about affairs incongruent to reality. If one can reason about beliefs that are false, then one understands belief. Hence, Wimmer and Perner (1983), and Baron-Cohen, Leslie and Frith (1985) kickstarted decades of research into false-belief understanding. An example of a classic test is thus:

> A child watches a scene where an actor places an object in one of two opaque boxes. The actor leaves the scene. A second actor moves the toy from one box to the other. The first actor returns, then the child is asked where the actor will look for their toy. To pass the test, one must understand that the actor has a false belief about the location of the toy; hence the child must intimate that it is the first (now empty) box that the actor will search in.

A focus of Baron-Cohen et al.'s work was also to demonstrate that autistic individuals struggle to pass these tests, suggesting that one of the defining characteristics of autism is non-neurotypical theory of mind. However, whilst psychologists were concerned with the ontogeny of theory of mind, philosophers were concerned with how one might go about attributing mental states to others. The positions considered most plausible were that either people generated and adhered to some hypothetical theoretical principles about mental states, something like developing a scientific theory about others' minds, or that people simulated others' minds and thought processes in order to understand how they think. Known as 'theory-theory', (for example in Gopnik and Wellman, 1994), and 'simulation theory', (such as in Gordon, 1986), over time the trend has been to converge on a hybrid with an emphasis on either theory (Nichols and Stich 2003) or simulation (Goldman 2006).[4] Due to the distinction between theory and simulation, 'mindreading' is now the preferable term over theory of mind, as the 'theory' in 'theory

[4] However, the discussion of simulation goes back to Hume (Hume in: Gordon 1995, p. 727).

of mind' might be seen as theory-theory endorsement. More will be said about theory and simulation for character traits in Chapters one, three, and four.

2.1.1 Implicit vs. Explicit Testing Methodology

It was held (based on a meta-analysis by Wellman, Cross and Watson 2001) that the ability to reason using knowledge of false beliefs developed at the age of four. However, since the 2000s, a growing body of evidence suggests that children demonstrate false belief understanding well before this age—even six-months-old in some cases (Southgate and Vernetti 2014). Bloom and German (2000) originally noted that when false belief tasks relied less on language—such as requiring pointing as a response rather than constructing sentences—or the language in the task less complex, children could pass tests at three-years-old instead of four. Surmising that language might be a bottleneck, Onishi and Baillargeon (2005) constructed false-belief tasks that did not require explicit language-involved questioning (and the response of the subject). Their test and subsequent replications and expansions—for example by He, Bolz and Baillargeon (2012); Knudsen and Liszkowski (2012); Scott and Baillargeon (2009); Senju et al. (2011); and Träuble, Marinović and Pauen (2010)—employed implicit methods like 'gaze-time tracking', 'violation of expectation', and 'anticipatory looking' paradigms. Such methods, broadly, are those in which the looking-behaviour of subjects are measured to infer anticipation, surprise, nonchalance, interest, et cetera. One might infer that an infant is surprised by noting the eye-widening, apparent attention, and focus on events when they are incongruent with what the infant could be expected to know.

Such methodology was met with scepticism, for example by Haith (1998), who maintained that to attribute knowledge of concepts to children on the basis of such behavioural prompts as eye-movement was absurd. However, this debate on looking paradigms was somewhat resolved prior to Onishi and Baillargeon's 2005 paper. The implicit methodology was challenged in response to tests for the ontogeny of 'object permanence', the understanding that things continue to exist when they are not observed (Baillargeon, Spelke and Wasserman 1985; Baillargeon 1987); subsequent researchers built specific conditions into their experiments to control for non-conceptual and mere behavioural interpretations that objectors were suggesting were responsible for gaze data. Examples of updated methodology can be found in Wang, Baillargeon and Brueckner (2004).[5] As such, looking-paradigm objections are becoming rarer, whilst implicit testing has flourished. Reviews by Baillargeon et al. (2015), and Baillargeon, Scott and Bian (2016), describe dozens of tests purporting to demonstrate that infants have surprisingly flexible false belief reasoning skills.

The research impact of the 2005 study with implicit methodology sparked a new debate, for a puzzle had arisen: before four-years-old, children can both *pass* (implicit) and *fail* (explicit) mindreading tests—how should that be understood? Attempts to account for this puzzle of mindreading come in several forms, mostly by constructing accounts of the mindreading system(s).

2.1.2 The Puzzle of Mindreading: Nativism or Empiricism Regarding Conceptual Development?

Studies of false belief understanding are essentially about when and how people come to acquire such concepts; hence, debates over the puzzle of mindreading are directly influenced by one's philosophical position on the acquisition of concepts. On one end of the spectrum, those with sympathies to 'nativism' regarding concepts are likely to attribute 'richer' conceptual interpretations of behaviour to infants, given the position of nativists that people can be born with, or have cognitive machinery dedicated to bootstrapping the acquisition of, particular concepts. Furthermore, for nativists, developmental milestones will usually be couched in the development of other cognitive abilities besides the acquisition of concepts (supposing that the concepts are already there, and that infants may have trouble accessing and using them whilst engaging in other cognitively taxing tasks). Those with more 'constructivist' sympathies generally make accounts that emphasise the learning mechanisms and subsequent *construction* of mental state concepts, as well as other cognitive developments; for example, in learning to develop hypotheses about the world that we can learn from (Gopnik and Wellman, 1994). Essentially, the difference is between cognitive machinery dedicated to helping us use the concepts that come online very early in cognitive development (nativist), and machinery dedicated to the construction and development of our concepts (constructivist).

Whilst we can conceive of the following answers to the puzzle of mindreading in terms of varieties of nativism vs. constructivism about concepts, we can also conceive of them

[5] See also Munakata (2000), who summarises weaknesses of objections to the implicit methodology.

as endorsing specific kinds of cognitive architecture. Whilst not exclusively endorsed by nativists, those with nativist sympathies tend to endorse a level of *modularity* about the mind, such that the mind is made of many cognitive modules that have been adapted for tasks and do not communicate significantly with modules that perform substantially different tasks. We can take the ‘massive modularity’ thesis (Carruthers 2006) as one end of the modularity spectrum. On the other end, more constructivist approaches may align with those approaches which take more domain-general mechanisms to be responsible for much of our cognition, where some take more holistic approaches to the mind’s structure—they tend to explain mindreading such that cognitive processes occurring throughout the mind broadcast activity to each other, such as Anderson (2014).

Putting positions on concepts and cognitive architectures together, we see that, for example, Heyes (2014) takes a heavily constructivist and non-modular approach (to learning, at least), arguing that behavioural reasoning and heuristics explain away the nativist-coded implicit testing data. Alternatively, Baillargeon, Scott, and Bian (2016), and Carruthers' (2013) nativist and modular approach maintains that the puzzle of mindreading is accounted for by cognitive bottlenecks that prevent infants from applying (through language) their already latent concepts of belief. Furthermore, Apperly and Butterfill (2009), and Butterfill and Apperly (2013), also held that children reason through false-belief tasks using mental concepts and not merely by parsing behaviour, but they maintain that the mindreading module is split in two. For them, ‘system 1’ uses ‘minimal’ concepts to achieve basic mindreading in limited scenarios, where ‘system 2’ utilises the full concept of belief to mindread fully; system 1 develops in infancy, system 2 develops at around four-years-old.[6] This sort of view takes more of a middle ground between nativism and empiricism, whilst still retaining modularity. Finally, others have suggested more holistic and less modular approaches to mindreading, in which multiple cognitive systems contribute to our learning and mindreading ability, such as (Christensen and Michael 2016).

What I have provided here is a history of philosophers and psychologists debating the structure of our mindreading system(s) in order to account for *how* we mindread. However, *my* focus is not on accounts of concept acquisition, nor questions of modularity, though there will be allusions to these matters throughout the thesis. For

[6] See also: Kahneman (2011).

example, it will become clear that I am not a nativist regarding character reading, despite my nativist sympathies regarding mindreading. Furthermore, it is doubtful that I can endorse modularity to any great extent. This is because of the amount of information, which comes from many different cognitive systems, that appears to be functionally useful in producing a trait attribution. This is elaborated upon in Chapter 3.

What ought to be noted is that this history of mindreading scholarship has thus far has been situated within a 'cognitivist' paradigm; that is, it assumes that the fundamental tenets of cognitive science are sound. Specifically, it assumes the soundness of the claim that the mind manipulates representational content via computation.[7] Of course, this invites alternative paradigm challenges, in this case from '4E' theorists, whose plurality of views broadly claim that the mind is in some sense 'Embodied, Embedded, Enactive, and Extended' (hence 4E).

2.2 '4E' Cognition and its Critics

4E positions generally claim that our cognition should be conceived of with relation to the phenomenology of (and dependence on) the body, and with relation to action. The views range from accentuating this view within a cognitivist paradigm alongside representations, to focussing on 4E instead of representations, to scepticism about representations and even content generally (Newen, Bruin, and Gallagher 2020). Regarding mindreading, this constitutes a major attempt to reframe the conversation in terms of the dynamicity of social interaction and the importance of non-mindreading socio-cultural aspects of folk psychology. For example, Daniel Hutto constructed a socio-cultural account of folk-psychological 'competence' that emphasised the learner's practical experience in engaging in and learning to construct/replicate social narratives (Hutto 2004, 2008). Similarly, Gallagher (2001, 2005) and Ratcliffe (2007), argued that social understanding is inextricable from situation in the social world—they reject the very idea that we commonly apply mental concepts in order to understand each other, suggesting that it is a sort of philosophical artefact abstracted from how we actually engage socially.[8]

[7] See also Shea (2013), on naturalising representational content *as* informational processing from the perspective of the cognitive sciences.

[8] For example, it appears that Hutto rejects the puzzle of mindreading—for him, mindreading develops by four-years-old, and Onishi and Baillargeon's data can be explained away by reference to behavioural associations, given that he rejects that infants represent beliefs, or even have beliefs with representational content, until they have acquired language (Hutto 2008, pp. 46–48).

As with any alternative paradigm, 4E approaches to mindreading have received criticism (Spaulding 2010a, 2010b, 2015; Michael 2011; Lavelle 2012; Bohl and Gangopadhyay 2014; Matthen 2014; Gangopadhyay and Miyahara 2015; Miłkowski 2015; Wheeler 2017; Herschbach 2018). The main thrust of the variety of criticisms is that 4E alternatives to cognition may highlight some real philosophical problems, usually about the nature of mental content, but their alternatives are shallow in explanatory power compared to conceiving of the mind as a representational machine.[9] For example, though he wrote before the enactive emergence, Pylyshyn (1980) presented a detailed picture of the positive reasons for conceiving of the mind as computational and representational. Furthermore, embodied approaches—such as focussing on cognition in relation to action—are considered by some to be include a 'tacit behaviourism' (where Skinner's behaviourism was famously critiqued by Chomsky, 1959), given that enactivists like Alva Noë would claim that the concept of a pain *just is* the concept of a state that makes you say 'ouch' (Block 2005, p. 262). Finally, a refocussing of perspective onto phenomenology may be unwarranted, given the many unconscious cognitive processes involved in mindreading; the fact that mindreading accounts will not always match our phenomenology of interacting with others does not mean that those mindreading accounts must be missing something (Spaulding 2015, though arguably the sentiment originates with Churchland 1981).

Whilst 4E challenges to mindreading have not revolutionised our understanding of mindreading, I do take such positions to have highlighted an area of neglect that warranted (and has now begun to receive) mainstream attention. This is the view that there are many other ways to understand others (and that there is more involved) than strictly mindreading. This recognises, particularly, that folk psychology requires research into the dynamics of social interaction, in addition to philosophising on the results of false-belief tests given in socially sterile circumstances.[10]

2.3 Folk-Psychological Pluralism

[9] Cognitive scientists are not unaware of such criticisms, and see them as an ongoing, though not paradigm-shifting, issue—see for example Piccinini (2015) on computation without representations, and Miłkowski (2013, chap. 4) on the role that representational content can play in computational cognitive mechanisms.

[10] This is not to say that this view was invisible to those working on mindreading—rather, that the focus was merely on understanding the results of false-belief tests. Eventually, such a focus was going to wane in favour of the next flavour of research.

Although I work within a cognitivist paradigm, I take the resulting 'pluralist' accounts (and the pluralism that I endorse) to be spiritual successors to the 4E challenge. Pluralism about folk psychology (Andrews 2008, 2012; Fiebich and Coltheart 2015; Fiebich 2019; Andrews, Spaulding, and Westra 2021) emphasises that we often understand others not just by attributing mental states but also by attributing character traits, by applying stereotypes, by creating and applying heuristics about behaviour, by recognising and responding to in-groups and out-groups, and by adhering to social schemas, et cetera. This is not necessarily a controversial nor ground-breaking statement, but it acts as a call to research—pluralists claim, much like 4E theorists, that mindreading is not even the *main* way in which we understand others. We might see evidence of this in practical reality, with the prominence of character and stereotype attributions relied upon in mainstream political discourse (Curry 2021, pp. 161–162). Mindreading, pluralists allege, instead of the main way in which we understand others, is an activity that we do when things go wrong, when the stereotypes do not match reality, when scripts are violated, when people act out of character, and that it is generally dependent on our social goals (Spaulding 2018b, pp. 6–7).

Even then, mindreading is often an inaccurate undertaking (Spaulding 2016). Given that there are many other ways of understanding others, and that the supposed main tool of folk psychology is often inaccurate, pluralists hold that we would be better off if we reject the 'primacy' of mindreading and the false-belief task, focussing on what it is in *everyday* social interactions that allows us to understand others.[11] This is also something of a departure from developmental psychology towards social psychology, as this call to research focusses on adult social interactions; this is not a rejection of the current literature, though, so much as it is a suggestion of the next topic of conversation.

However, one might accept the pluralist call to research without accepting a demotion of mindreading—what has generated debate is the question of the extent to which

[11] Whilst mentioned for posterity, the accuracy of our folk-psychological skills is a side-issue. As we might expect of attempts to understand those who are not ourselves, Andrews (2008) and Spaulding (2016) have made the case that mindreading is often inaccurate, whilst Westra (2018, 2020) argued that character trait attribution is often inaccurate. William Ickes, who discusses 'empathic accuracy' as the extent to which we are correct about reading the *specific* thoughts and feelings of others, notes that even married couples only get it right about 35% of the time; in his whole career he saw highs of only around 60% (Ickes 2011, p. 201; Maibom 2017, chap. 31). Despite these data, as a general observation, humans seem to be fairly successful in our understanding of others on a day-to-day basis, at least to the extent that it does not hamper *fluid* and *mutually useful* social interactions. It is currently unclear to what extent research on the parameters of accuracy of folk-psychological judgments pertains to claims about the primacy of mindreading, or other issues; hence, this issue will be shelved in this dissertation.

mindreading is the *main* aspect of folk psychology, and whether mindreading can be separated from other tools in our folk-psychological toolkit.[12] For example, whilst Andrews (2012, p. 102) tentatively suggests that non-human animals might use character in socially cognitive judgments (despite not having mindreading capacities), Spaulding (2018a) argues that social interactions are too complex to separate mindreading from other tools; trait attributions affect mindreading, and mindreading affects trait attributions. This view is also shared by Westra (2018), who conceives of trait attribution as being a part of a mindreading hierarchy of informational processing.

I perceive the debate over the primacy of mindreading and the legitimacy of pluralism as beginning with character traits (and so character reading). Pluralists have often referred to character traits as a means to understand others, helpfully framing it like mindreading, noting that we attribute character traits to others in order to predict or explain their behaviour (Andrews 2008, pp. 16–26). Hence, it is prudent to examine and contribute to this work now. That said, whilst mindreading has a history that has led to the pluralist challenge, character traits also have a philosophical and psychological history, which needs reviewing before I present my thesis on character traits in the context of mindreading.

3. Contextualising Character Research

3.1 Traits and the Philosophy of Dispositions

The first two questions that motivate philosophical and psychological research on traits are the following: What are character traits? Are they real? Regarding the former question in philosophy, traits have generally been conceived cross-disciplinarily as dispositions to behave, or think, in certain ways. These dispositions are stable across time, though acting out of character is common and traits can change over the long term. However, traits do not seem to match all of the paradigmatic features of dispositions. Whilst dispositional properties need not manifest, like a glass being fragile despite never breaking, character traits *do* need to manifest to be possessed; we do not call people brave if they have never been brave in their life (in thought or action). Hampshire (1953) claims that to make a statement about a trait is to attribute a positive history of disposition manifestation: traits are true dispositions, and 'dispositions' like fragility are descriptions of a thing's causal properties. However, modern work on

[12] There is, however, an interesting question as to whether the views are compatible as it pertains to the nature of practical reasoning—this is discussed briefly in Chapter 5.

dispositions seems to settle on fragility nevertheless being a disposition (for example Mumford 2003; Hacker 2007; Vetter 2015). Still, there is a difference between fragility and bravery that needs explaining. Alvarez (2017) focussed on this distinction and tentatively suggested that traits be considered as 'tendencies', as this better captures the history of dispositional manifestations implicit in trait attributions. Chapter two, 'What is a character trait?', discusses what character traits are in detail.

3.2 Empirical Accounts of Traits: The Five-factor Model

In psychology, modern work on traits began in the 1930s. Traits tend to be defined as the dimensions of individual differences between people in showing particular patterns of thoughts and behaviour across time (McCrae and Costa 2003). Psychologists also recognise traits as dispositional and stable over time, where trait expression can be internal as well as behavioural (Allport 1931). Much modern trait work focusses on dimensions of difference accounted for by the 'five-factor model'.[13] The five-factor model is not an attempt to account for every trait, like bravery, cowardice, or generosity, et cetera. It was constructed out of research into what basic dimensions of difference people express (based on self-reports). Therefore, the five-factor model is constructed from the dimensions of our folk notion of traits, i.e., the pre-theoretical everyday understanding. It claims that the five basic dimensions of personality are: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Take *anxiousness*: this fits along dimensions of neuroticism, extraversion, and openness; *generosity* fits into dimensions of conscientiousness and agreeableness.

Despite wide adoption, the five-factor model has received some criticism. For example, some have argued that 'honesty' should constitute a sixth dimension (Ashton, Lee, and de Vries 2014). Others have provided alternative accounts, such as the 'temperament and character inventory' (Cloninger 1986), which takes the basic dimensions of character to be self-directedness, cooperativeness, and self-transcendence; it is arguable, though, that these dimensions collapse into five-factor terms (De Fruyt, Van De Wiele, and Van Heeringen 2000). Furthermore, there is the worry that personality psychology relies on the truth of the assumption that all of the relevant trait descriptions in our psychology have been sedimented into natural language, i.e., that our real psychological traits are accurately and fully described by our language (Cattell 1943). If this is not the case, then we may have to re-evaluate our claim that the five-factor model correlates

[13] See Goldberg (1993), for historical review.

with behaviour, as the five-factor model does not merely observe that five factors underlie our traits but further claims that they are in some sense explanatory.[14]

Despite these criticisms, the five-factor model remains the most prominent account of the underlying dimensions of traits and part of its strength comes from not claiming to make any predictions of individual behaviour based on trait assessment. However, it can be applied to wider trends in one's life, and has been shown to be quite accurate in those predictions (Matthews 2015). This review is mainly historical context for psychological research on traits. Here, I make note of why I am not committed to a critical importance of the five-factor model to philosophical work on character reading, and in 3.2.1 I point to the kinds of empirical data that I will be using in this dissertation.

This is because there is a worry about how personality data may be used. Wille, De Fruyt, and De Clercq (2013) found that the five-factor model—in relation to 'aberrant' traits like narcissism and anti-socialness—had the power to predict the outcomes of subjects' careers over fifteen years. Narcissistic anti-social people tended to have high-earning high-management level careers; neuroticism was a big predictor of current job satisfaction and stress after fifteen years. The study itself then suggested that their results could be useful for *human resources hiring practices* (p. 212). Research of this kind, therefore, can be categorised under the slogan of 'we can, but should we?'[15] This is an important question, but not one that I have space to argue for my reticence about; I will keep my account of character reading free of a necessary commitment to the five-factor model.

3.2.1 Empirical Accounts of Traits: Spontaneous Trait Inferences

There *are* data about traits that this dissertation will appeal to, though, which take the form of accounting for sub-personal attribution of trait concepts. Winter and Uleman (1984) first studied so-called 'spontaneous trait inferences'. In order to establish whether traits were inferred and stored in memory at the encoding of observed behaviour, Winter and Uleman had subjects read and remember sets of sentences designed to evoke trait attributions, then had them recall the content of those sentences later. The reasoning being, according to the 'encoding specificity principle' (Tulving and Thomson 1973), that if trait attributions had been made when behaviour was encoded,

[14] Though, see Fleeson and Jayawickreme (2015), who provide a *descriptive* component of traits with five-factor data, and an *explanatory* component given by social cognitive mechanisms.
[15] Arguably, we *cannot* give the unreliability of attributions based on such data.

then a memory retrieval cue of the same kind of information will exhibit higher successful recall. Hence, if a sentence was about someone having his neighbour round for dinner and the subject had encoded ‘friendly’, then the cue word ‘friendly’ would result in better memory recall of that sentence. Subsequent positive results from this study encouraged more research into spontaneous trait inferences; for example, Ham and Vonk (2003) obtained more evidence using similar methods. Furthermore, Fiedler and Schenck (2001) replicated and expanded in Fiedler et al. (2005) recorded spontaneous trait inferences made in response to pictorial representations and videos. The relevance of spontaneous trait inferences will be apparent in Chapters three and six.

3.3 Problems for Empirical Work on Traits: The Fundamental Attribution Error

The psychological identification of traits leads to the next question posed in 3.1: Are traits ‘real’? This question has two motivations that require slight diversions. The first is psychological evidence against traits in the form of correspondence bias, also known as the fundamental attribution error. This is invoked in classic debates over traits as virtues/vices. The second motivation is based on a general distrust of psychological scholarship (particularly social psychology) for two reasons: the discipline-wide ‘replication crisis’, and the scepticism of scholarship that makes universal, cross-cultural claims about human psychology (see 4.2).

The fundamental attribution error is that people tend to attribute what others do to internal character, as opposed to considering the situation as the relevant factor (Jones and Harris 1967). Imagine an interviewer who is asking difficult questions of an interviewee. People tend to attribute more intelligence to the interviewer, despite spectators knowing that the interviewer has answers to the questions. If (whilst driving) someone cuts you off, people tend to attribute traits like stupidity to the driver without considering that there may be a relevant situational factor accounting for such behaviour. A classic example of a situation trumping traits is that whether or not you are in a hurry is the biggest factor in whether you stop to help someone in need, even if you are a theology student hurrying to give a presentation on the ‘good Samaritan’ parable (Darley and Batson 1973). The question is that if situational factors are key, and if we mistake situational factors for traits, then do such folk notions of traits as virtues and vices even exist? Doris (1998), Harman (1999), and other ‘situationists’ argued that they are indeed illusions.

However, this does not entail that character traits do not exist. The situationist challenge alleges that we treat traits as predictors of behaviour when in fact situations are. Personality psychologists do not claim that traits are good predictors of behaviour in *specific* situations. Traits derive their explanatory power through predictions across time and situations; traits are excellent predictors of general trends (Stagner 1977, especially sections B and E).[16] A more comprehensive investigation on the situationism debate needs to be conducted in order to definitively establish the existence of traits, though a meta-analysis of tens of thousands of twins in personality studies ((Vukasović and Bratko, 2015)), produced a statistically significant result on the heritability of personality, which is at least some evidence in support of this thesis.[17] For now, we can end on the note argued for in Lamiell (2018), that whilst the complaint is about the consistency of trait manifestations between individuals, the evidence cited by situationists is only interpretable about differences between *populations*. Furthermore, the evidence that situationists use are not about differences between population members in general, but rather *in aggregate*. A 'general' fact established by statistical analysis demonstrates that this fact is common to all individuals, but an 'aggregate' fact has no such commonality to each individual population member. All that the situationist evidence can show, Lamiell argued, is that individuals were not equally consistent in their trait manifestations, which is to be expected.

3.3.1 Problems for Empirical Work on Traits: The Replication Crisis

Another issue, foreshadowed by the care I have taken to note replications of studies for particular psychological effects (such as spontaneous trait inferences), is the replication crisis. This is the recognition in academic psychology (particularly social psychology) that many works published since the 2000s have low-to-no replicability i.e., we are unable to replicate the original results; this was demonstrated in a meta-replication study (Open Science Collaboration 2015). Furthermore, many psychologists admitted in anonymous surveys that they flouted best practice, such as by reporting unpredicted results as predicted, or not reporting all of the measures for data collection (see John, Loewenstein, and Prelec 2012). Furthermore, one could generate ridiculous results and

[16] This is not to say that *no* psychologists are working on demonstrating how traits could be predictors in specific instances, but this is not the normal consideration—see Matthews (2018) for development on this.

[17] See also: Bouchard and Loehlin (2001, p. 258) for references to other twin studies on thousands of twin pairs.

be published—see Bem (2011), who reported that *future* events could influence *current* subjects' answers.

An increase in public worry has led many to be sceptical of psychological studies, even though a lack of replicability does not strictly falsify *all* of the results of an original study—see Earp and Trafimow (2015) for a discussion of how false and non-replicable results should affect our confidence in both the studies and wider psychology.[18] Though steps have been taken to improve the quality of studies since the crisis was brought to the public eye, structural issues in academia surrounding the constant pressure to publish and the lack of prestige associated with replicating others' work have not been addressed; thus, it is still best practice to cite replications where possible.

Regarding replication crisis worries for work on character traits, let us consider the five-factor model. The five-factor model has achieved prominence partly based on its repeated cross-cultural corroboratory results (McCrae et al. 1998) but also based on failures to provide better validation of other theories (McCrae and Costa 1987). Furthermore, traits have a degree of genetic heritability (McCrae and Costa 2003; Power and Pluess 2015; Bratko, Butković and Hlupić 2017). Given the plethora of studies on the five-factor model and its replicated results, due diligence has been performed on the five-factor model. Similarly, the studies I discuss in this dissertation received multiple citations for replications and expansions where possible; even in this literature review, this can be found when discussing the implicit testing methodology for false belief tests, for example.

We might identify a final objection to the reality of character traits. A cross-cultural basis and genetic heritability may evoke moral worries of the kind associated with the unscientific Myers–Briggs personality tests (Myers and Briggs 1962). Myers–Briggs tests try to categorise people based on personality, making judgments about how people typically behave and who they are as people. By contrast, the five-factor model is not generally for this purpose (cf. Goldberg, 1993, p. 32). Whilst there is an element of genetic heritability, psychologists do not make claims that our traits are innate or universal. Nor do they use the five-factor model to determine how we *will* act or how we *should* act (cf. Goldberg, 1993). Debate over whether character traits are 'real' by

[18] Similarly, one must not be taken in by the 'fallacy' fallacy, though the sheer scale of non-replicability in psychology was actionable here.

reference to scepticism induced by Myers–Briggs is not so much a legitimate debate as a common misunderstanding over what constitutes useable work in psychology.

4. Towards the Social Cognition of Traits

Having outlined some of the empirical work on character traits in psychology and briefly defending their study against the replication crisis and the situationist challenge, we can return to pluralism and the question of how information about character traits is processed by the brain—in essence, the social cognition of character. In particular, even if character traits are not real *per se*, our brains still use these concepts in some manner that warrants explaining. A question of the cognitive architecture of character reading is therefore not invalidated by concerns about the metaphysical reality of traits. On that score, the most notable account of this is Evan Westra's hierarchical predictive coding account. I will briefly outline what predictive coding is and the main thrust of Westra's account, but this dissertation is not a thesis that uses predictive coding in its explanations. I see my work as being compatible with both predictive coding generally and Westra's account, merely framed at a level of psychological reasoning where predictive coding architectures might be explanatory in enabling such reasoning. However, I do note the instances in which my work departs from Westra's.

4.1 Predictive Coding and the Architecture of Information Processing of Traits

Predictive coding is a theory imported from neuroscience.[19] The theory holds that the brain is constantly generating models of the environment, such that it can predict incoming sensory inputs; it then updates the models based on which inputs are successfully predicted and which are prediction errors—see Clark (2013). The upshot is that it purportedly provides a unifying account of perception and action by showing how information is processed in cortical hierarchies.

As a brief analogy, imagine playing the party game 'twenty questions'. In the game, someone writes something on a piece of paper, and you have twenty chances to guess what it is by asking 'yes or no' questions. Remarkably, people can be quite successful at this game without using the full twenty questions. The idea is that you have certain expectations about what is on the paper, and you update your expectations accordingly, in order to minimise uncertainty. "Is it an animal?" or "Is it a fictional character?" are

[19] Whilst neuroscience is undergoing something of a paradigm-shift as a result of predictive coding, predictive coding is also responsible for many successes in the programming of modern machine-learning. See Millidge, Seth and Buckley (2021).

common starts; thus, you begin the game with a certain set of expectations about what might be written down. In essence, you are not initially expecting them to write a differential equation, though plausibly you could adjust this expectation as you ask your questions. Your questions therefore become more precise depending on the previous input, eventually asking a final question that banishes uncertainty. Analogously, your brain is constantly playing twenty questions with what it expects from sensory inputs from the environment, based on what it has already received. The brain is so good at this game that it lets us both perceive and react quickly, whilst learning about our environment efficiently.[20]

With predictive coding in mind, Westra built an account of mindreading. The basic idea is that at the relative top of a decision-making hierarchy, we can use character information to help infer the probability of certain mental states of targets, then use our inferred mental states of others as predictors of their behaviour. In turn, their actual behaviour may generate prediction errors that make us adjust our predictions of their mental states (and then potentially higher up the hierarchy to their character traits). We will see that my account of character in cognition is not incompatible with a predictive coding framework; it is, however, framed at a level of explanation that includes conscious reasoning and is thus more of an account of the structure of character-based reasoning, rather than a neuropsychological account of character-information processing.

My account departs from Westra's, though, in that my conceptualisation of mindreading differs. As will be detailed in Chapter one, my conception of mindreading is narrower than Westra's; hence, character reading—whilst dependent on mindreading both ontogenetically and phylogenetically—is not itself a part of mindreading. My account thus differs from Westra's framework *prima facie* on a terminological issue (given that Westra is sympathetic to pluralist approaches to folk psychology), but the difference bears fruit in Chapters five and six in detailing the relationship between character reading and mindreading.

[20] Strictly speaking, the twenty questions analogy is more akin to describing Bayesian inference than predictive coding *per se*. See Berger (1988) for an overview of Bayesian inference in statistical theory, and Etz and Vandekerckhove (2018) for its relevance to psychology. That said, Bayesian inference as a statistical equation for updating models of uncertainty is a large part of predictive coding; I see further detail on predictive coding as extraneous and potentially confusing (for the purposes of this dissertation); hence, we will make-do with this analogy.

If not produced on an architecture of predictive coding, what will my account of character in social cognition involve? Essentially, my architectural commitments are related to the structure of processing in acts of reasoning, both conscious and unconscious (where the outputs of reasoning could be considered within a framework of hierarchical processing, if necessary). I focus heavily on the issues of applying *theory* and *simulation* to character reading; whilst I remain (officially) neutral on the status of mindreading, character trait attribution and subsequent reasoning can be situated within a constructionist paradigm that allows for a pluralistic approach to folk psychology whilst, crucially, retaining the primacy of mindreading. Themes of the effect of culture on social cognition (and character) should also be apparent throughout this dissertation, which constitutes the final contextualisation before I get into the details.

4.2 WEIRD Psychology and the Impact of Culture on Social Cognition

A final point on the psychological scholarship of character traits is the impact of doing psychology on mainly Western, Educated, Industrialised, Rich and Democratic (WEIRD) populations (Henrich, Heine and Norenzayan 2010a, 2010b; Henrich 2020). Essentially, WEIRD individuals have historically been the main subject of academic psychology. This has resulted in some erroneous conclusions as to what *human* psychology is like. For example, Henrich, Heine, and Norenzayan (2010b, sec. 3.1) point out that early anthropological work such as that of Segall, Campbell, and Herskovits (1966) has shown that different cultures can perceive visual illusions (such as the Müller-Lyer illusion) differently.[21] However, such cultural considerations are rarely central to psychological theses.

Cultural differences can be found with character reading and are thus emblematic of pluralistic folk psychology as a conglomeration of socially cognitive skills that are developed and practiced in varying ways. For example, American adults often make more trait-based inferences about others compared to inferences based on situations, as is the case for some Indian Hindus (Miller 1984).

It is thus important for work on character reading to be cognizant of cultural differences, and for such accounts to provide avenues of explanation for such differences. Lavelle (2021) argues that accounts in social cognition should include

[21] Segall et al.'s hypothesis was that the 'carpentered' nature of rich Western environments biased development of the early-visual system, where Ahluwalia (1978) expanded this finding and provided evidence that cross-cultural results were not the result of genetic differences across cultures.

explanations of how cultural differences affect psychological reasoning, as 'guiding features' for those accounts in understanding. I plan to fulfil that aim for character reading in social cognition in the acceptance of pluralism about folk psychology (with caveats), in developing the pluralistic methods of reasoning presented in Chapter four, and in the development of trait attributions as per cultural evolution in Chapter six.

As such, in this introductory chapter I have detailed the rich psychological and philosophical history of the study of mindreading-centred social cognition. This has culminated in a drive to explain the wider context and roles of certain skills for understanding others that go beyond merely attributing thoughts and feelings to them, and to explain their relationship to mindreading in a way that is sensitive to cultural considerations. I focus on character trait attribution and reasoning as a prime example of such an alternative socially cognitive practice; the final section gives an overview of each chapter of this dissertation.

5. Chapter Summaries

5.1 Chapter One

In this chapter, I will detail key terms: cognition, social cognition, folk psychology, mindreading, theory, and simulation. I will offer context and detail, as these are practical as well as stipulative definitions. I will argue that it is most useful to conceive of mindreading as 'mental-state attribution', as opposed to a broader folk-psychological understanding. I will argue that on the understanding of theories as being *theory-making* according to the 'special vocabulary' that constitutes them, one can constrain the current socially cognitive literature into theory-based accounts, simulation-based accounts, and hybrids; radical and alternative constructions collapse into theory, in my view.

5.2 Chapter Two

In this chapter, I will give my account of what a character trait *is*, clarify key terms within the definition, and defend the account from issues stemming from the metaphysics of dispositions, such as objections about their causal efficacy. Furthermore, I will argue that whilst dispositions are important in understanding traits, character traits are more than mere dispositions because they include some history of disposition manifestation; as such, I claim that character traits are 'tendencies', and I will defend this definition.

5.3 Chapter Three

In this chapter I will motivate the study of simulation as a process in character reading, beginning this investigation with a comparison to existing work on empirical accounts of simulation in mindreading. I will discuss Goldman (2006) on simulation in sub-personal mindreading, showing that we can rule out simulationist accounts of sub-personal trait *attribution*. Nevertheless, a hybrid account of character *reading* is still possible. Furthermore, it is likely that simulation in trait attribution occurs as part of the identification of particular trait-relevant features that provide information alongside theoretical knowledge to form trait attributions.

5.4 Chapter Four

Developing on from Chapter three, in this chapter I will give my hybrid account of character reading. I will argue against a 'theory-theorist' account of character reading; then, I will argue for simulation in character trait attribution and reasoning. I will argue that simulation is mechanically present in trait *attribution* when we recall information from memory in order to construct imaginings of others' trait-relevant behaviour, and that simulation is present in character *reading* when we attempt to become another for the purposes of understanding them; crucially, I argue that has previously been mistaken for mindreading.

5.5 Chapter Five

In this chapter, I will detail the conceptual and explanatory relation of 'dependence' between mindreading and character reading, such that the trait concepts required for trait attributions and trait reasoning are dependent on a particular self-knowledge of our own mental states. With this, I will argue against the current pluralist interpretation that character reading is entirely independent of mindreading, and further argue that my pluralist account is preferable because it allows us to explain why explanations of others' behaviours by reference to their traits are explanatorily satisfactory in a way that current pluralists fail to capture. My account is pluralist in that it considers the importance of varying methods of social cognition, but it does not reject the primacy of mindreading.

5.6 Chapter Six

Given the similarity and dependence of character reading on mindreading, in this final chapter I will fulfil the promise of accounting for character reading in social cognition in relation to variance across cultures. I will explain and defend the existence of so-called ‘cognitive gadgets’: cognitive mechanisms which are the product of cultural, rather than genetic, evolution. I will argue the positive case for trait attribution being a cognitive gadget, considering, for example, that the mechanisms which allow spontaneous trait inferences need to adapt faster across environments than genetic evolution would allow. As such, this chapter doubles as an account of the ontological dependence of character reading on mindreading. I will finally defend this constructionist account against a plausible nativist alternative.

Cognising Social Cognition

1. Introduction

Many of the terms introduced in the introductory chapter require elaboration. In order to make claims about character reading, the research must be contextualized within the broad field of social cognition. Within social cognition, character reading must be situated relative to the study of folk psychology and mindreading. Mindreading and character reading function according to theory-laden and simulative processes. This chapter stipulates and contextualises my use of these terms. I move from the broadest topics to the narrowest: social cognition, folk psychology, mindreading, then theory and simulation.

Section two details social cognition, folk psychology, and mindreading. Section three gives detail on theory-theory, then provides a brief account of what makes a theory; the upshot is that this notion of theory constrains the dialectic to a dichotomy between theory and simulation only. Section four gives two notions of simulation and settles on *process* simulation for this dissertation. Finally, I consider and respond to the objection that simulation collapses into theory.

2. Social cognition, Folk Psychology, and Mindreading

2.1 Social cognition

Social cognition regards those thinking processes that allow for action *in relation to* others and interaction *with* others; I call such action 'social action'. This includes the storing/retrieval of information and its manipulation, for example through reasoning. Whilst there are norms to the term's use, there is no standardised definition. In so far as it *is* defined, it is a catch-all term used to refer to those cognitive processes which are studied when we research how creatures interact with each other. For example, in a recent textbook on the philosophy of the social mind, social cognition is defined as:

> Our ability to interact socially with other people. It is the broadest umbrella term used in the book, and all the theories under discussion aim to elucidate some aspect of our social cognition (Lavelle 2019, p. 4).

Because of its use as a broad term, it often becomes hard to identify its boundaries. It is fine for social cognition to be defined as above in a textbook about *human* minds, but social cognition is not necessarily anthropocentric. We might compare complex human reasoning with the study of social cognition in ants (for example as in Feinerman and

Korman 2017); these both study the same phenomena, but ant behaviour is clearly nothing like the outputs of human reasoning. Furthermore, whilst it is only implied or assumed, social cognition is treated as the study of how we *understand* others, not just how we *interact*. Below, I will stipulate two senses of cognition, one commonly used in cognitive science, and one used in philosophy. Doing so will provide some boundaries between elements on the spectrum of social *cognition*, whilst allowing me to clarify that the aspect of social cognition I am mainly concerned with in this dissertation is the 'conservative' rather than 'liberal' notion of cognition.

2.1.1 *Conservative and Liberal Cognition*

Heyes, in Bayne et al. (2019), notes that there are two notions of cognition which are generally employed across the cognitive sciences and philosophy. Cognition conservatives treat cognition as when "a cognitive process involves reasoning. It operates on propositions (sentence-like mental representations), involves beliefs, desires and other intentional mental states, and is typically available to conscious awareness" (p.r611). Perhaps this definition is *too* conservative. There are sceptical challenges to belief and desire's role in cognition, classically by Churchland (1981), who is optimistic that an adequate neuroscience will eliminate the need for beliefs and desires in cognitive explanations. Such a debate is important but should not be had within the context of a broad categorisation of cognition. I stipulate a neutral position for conservative cognition such that conservative cognition merely involves reasoning and operates over that which can be described by propositions. As such, conservative cognition requires a mind and capacity for reason. Humans clearly have this, maybe some animals do, but ants do not. Traditionally (though not exclusively), this is the sense of cognition that philosophers are interested in.

Cognition liberals, on the other hand, claim that "when we say a process is cognitive, we mean that it handles information in an adaptive way and can be modelled usefully as a form of computation" (Heyes, ibid). Modellable information processing (if it is adaptive) counts as cognition, even if that information is not explicitly reasoned about by the entity. An example would be the behaviours of swarms of bees in their co-ordinated picking of nesting sites (Passino, Seeley, and Visscher 2008). Liberal cognition is traditionally (though not exclusively) the interest of cognitive science; for cognition liberals, *social* cognition would include the social actions of ants.

Given these definitions, note that conservative cognition is a way of characterising stipulations on liberal cognition (ostensibly to discuss cognition at the level of psychology, rather than purely about information passed between neurons).[22] A conscious reasoning process will be (in theory) computationally modellable; reasoning about concepts is conservative cognition, but it is also a particularly demanding and specific form of liberal cognition. For example, an ant leading fellow ants to food, and a human doing the same, are both cases of cognition in relation to social action. These may be described in the liberal sense. However, in the human case we must stipulate additional capacities (for reasoning) to fully *capture* what is involved in the understanding invoked by the cognitive processing. Whilst bees picking nesting sites may be communicating and adaptively responding to information from the environment, from a conservative standpoint they would not be interesting targets of study because in order to have concepts of the informational content they manipulate, bees would have to recognise such information as having particular content (which they appear not to).[23]

In this dissertation I will be using the conservative notion of cognition, given that I discuss reasoning with the relatively complex concepts of character traits.[24] That said, I noted that conservative cognition is a certain way of characterising complex liberal cognition because some of the empirical data I draw on (examples throughout) have been computationally modelled. With knowledge of what I mean by social cognition, we can turn to folk psychology.

2.2 Folk Psychology

Much is referred to as folk psychology. To quote Ravenscroft (2016), there are at least three commonly used senses:

> 1. As a particular set of cognitive capacities which include—but are not exhausted by—the capacities to predict and explain behavior.
>
> 2. As a theory of behavior represented in the brain.
>
> 3. As a psychological theory constituted by the platitudes about the mind ordinary people are inclined to endorse.

[22] Liberal cognition might also occur in glial cells, alongside neurons (Peteron 2021).

[23] That said, there is some interesting research suggesting that bees may possess something like a rudimentary concept of time, given their ability to feed themselves at a particular time despite their circadian rhythms being disrupted by living in a salt mine, or experiencing time-zone change (Beling, 1929; Wahl, 1932; Renner, 1960).

[24] I discuss the complexity of the concept in Chapter five.

Here, I stipulate my usage, then explain how it relates to these three common notions.

I take folk psychology to be the collection of strategies we employ in understanding others, which are given by a set of socially cognitive capacities. These capacities allow us to engage in, respond to, and reason about the behaviour and cognition of others as if these were governed by implicit rules of what we take to be 'common-sense' about psychology (for example, that people *want* things and *know* things). As such, folk psychologies will differ in terms of the knowledge of (and application of) strategies across cultures, but a capacity for understanding others according to common-sense psychology is universal if one has a folk psychology. One can use these socially cognitive capacities to learn concepts of belief and desire, and to apply them even if those particular concepts are alien to that culture (or if it turned out that, strictly speaking, nobody ever wants or knows anything).[25] Individual cognitive mechanisms that provide the capacity for understanding others, according to specific strategies, may be characterised as folk-psychological, though how those mechanisms function need not be described according to common-sense psychology.

Folk psychology so described does justice to Ravenscroft's noted senses, whilst being preferable, for several reasons. Firstly, my definition allows a clearer distinction, compared to (1), between folk psychology and those cognitive mechanisms that grant such strategies. Secondly, an implicit theory about why others behave as they do is not committed (as the spirit of 2 is) to the claim that folk psychology *just is* a theory of behaviour. This is useful because there are reasons for thinking that folk-psychological competence is underpinned not only by implicit theory but also by simulative practices. Folk psychology (in my view) allows for these kinds of explanation of implementation to co-exist to any extent that they might. The third sense (3), pertains to David Lewis' work on the meaning of theoretical terms (Lewis 1972). More will be said in 3.1.2, but the point is that 'platitudes about the mind ordinary people are inclined to endorse' is included, in my case, as a body of information contributing towards our common-sense psychology. My definition allows compatibility with this thought that the contents of

[25] See Dewhurst (2017, chap. 2) for an overview of differences in expressions of folk psychologies across cultures through lenses of anthropology and comparative linguistics. Whilst Dewhurst rejects the 'folk-psychological universality' claim about explicit reports of others' psychology across cultures, he and I differ on the nature on the complete universality of folk-psychological mechanisms. Dewhurst uses evidence from implicit false belief tasks to argue for the plausibility of the claim that basic social cognitive mechanisms are universal, whereas we will see in Chapter six that I do not think this holds for all socially cognitive mechanisms: in some cases, a socially cognitive capacity's cognitive mechanisms are products of cultural evolution, rather than universal genetic evolution.

folk psychology are statements about psychology that *the folk* would endorse, whilst making room for the pluralistic position that highlights the strategies we employ to understand others.

2.3 Mindreading

The notion of mindreading I use falls into one popular camp of understanding, in which I seek to maintain the perceived importance of 'mindreading' as a key socially cognitive skill. I take there to be two senses in which mindreading is meant. One is 'folk-psychological', pertaining either to propositional attitude attributional practices themselves, or to a wide sense of folk-psychological skills that go beyond propositional attitude attribution. The sense of mindreading that I endorse treats mindreading as a particular ability (rather than as a set of practices) that is involved in enabling some (but not all) folk-psychological strategies: mindreading is the ability to attribute mental states to others. I shall detail both and justify my choice.

2.3.1 Mindreading as Folk-Psychological

In the folk-psychological construal, mindreading is considered to be a tool for understanding others by reasoning about the propositional attitudes attributed to others (and in one account to oneself as well—Carruthers et al., 2012). An attitude towards a proposition generally takes the form of a belief or desire. X believes that Y (and so may Z). X desires Y (and so may Z). In this view of mindreading, mindreading is *all about* attributing beliefs and desires.[26] The ability to have an attitude towards a proposition entail being a minded being—some agent rather than some object. Mindreading in this sense, then, refers to the strategy of attributing propositional attitudes, underpinned by an understanding that others have minds and that those minds have ongoing beliefs and desires.

As I have already noted, recent scholarship has focussed on how folk psychology either cannot be centred around mindreading (in the propositional attitude sense), and/or has suggested that mindreading study should focus on much more than beliefs and desires. For example, Spaulding (2018) notes that:

[26] Shannon Spaulding notes that it is often narrower than this. Often, only beliefs are the main subject of papers in mindreading research. She outlines a historical case for this in Spaulding (2018b, p. 2), but see Merricks (2009) and Lycan (2012) for the view that whilst belief is a propositional attitude, 'desires' are not. C.f. Buchanan (2012) for scepticism that even belief is a propositional attitude.

> left out of the standard mindreading story are the ways in which situational context influences in-group dynamics, which behavioral patterns and stereotypes are salient, the personality trait inferences we make in various contexts, and the biases that shape how we interpret a social interaction (p.25).

Regarding the folk-psychological understanding of mindreading, we can see why confusion and crosstalk occur between mindreading and folk psychology, given some people's conceptions of these terms in relation to propositional attitudes. Indeed, talk of propositional attitude attribution begins its history with folk psychology. Churchland (1979, 1981) explicitly describes folk psychology as involving propositional attitudes, and as Dewhurst (2017) notes: "Fodor also makes an explicit claim about the universality of propositional attitude psychology (or at least beliefs and desires), even going so far as to deny that there are any conceivable alternatives" (p.29). Fodor's was a restrictive view of folk psychology, but the point is to highlight how this historical prevalence of propositional attitudes has affected how some authors conceptualise the term 'mindreading' and subsequent challenges to its study.

What is jarring about the folk-psychological propositional attitude sense of mindreading is that, as a strategy, it would be rare in its explicit form. Consciously reasoning about others' beliefs and desires is comparatively rare in our day-to-day lives. I may witness someone drinking a glass of water and I may understand why they are doing this without explicitly attributing to them the belief that there is water in the glass, the belief that it is potable, the belief that drinking it would quench a thirst, et cetera. But then why has mindreading been the main focus of social cognition work for forty years? In other words, why do scholars want to take mindreading seriously?

2.3.2 Mindreading as Mental-State Attribution

The historical context of the literature (section 2 in the introductory chapter) has played a part in shaping mindreading as a key socially cognitive ability. However, whilst it is true that research into our understanding of the mental states of others has focussed on propositional attitudes, it seems disingenuous to attribute the importance of mindreading solely to historical context.

Indeed, it does not seem to me that those who work on mindreading conceptualise mindreading as merely belief/desire attribution. I suppose that most in the mindreading canon treat mindreading as *mental-state attribution*, and it is for this reason that mindreading takes the 'prime' spot in discussions of folk psychology. In this sense, mindreading is incredibly important. Regardless of whether we are specifically attributing beliefs and desires, we are (at the very least) attributing *mental properties*. This sense of mindreading avoids the propositional attitude-based critique; it does not specify *which* mental properties are attributed to others. This notion explains why we care about mindreading, and clearly delineates it from similar socially cognitive objects of study.

Here, I will also attempt to have my cake and eat it too. I treat mindreading as mental-state attribution specifically, but often the literature will discuss it in the sense of understanding others according to their mental states. Hence, my talk of *character reading* fits this structure as well. Character reading may be split into its specific sense of character trait attribution and then a sense of understanding others according to the traits we have attributed to them.

In sum, mindreading is mental-state attribution, which is an important socially cognitive capacity because it allows us to consider the contents of others' minds beyond the interpretation of their behaviour. It underpins many folk-psychological strategies, such as understanding others by reasoning about their beliefs and desires. This marks my departure from a strict endorsement of pluralistic folk psychology thus far because the importance of mindreading is retained under this definition. Having given stipulations on my terminology for social cognition, folk psychology, and mindreading, we can now move on to the terminology that underpins my discussion of character reading: theory and simulation.

3. Theory, Theory-Theory, a Theory of Theory-making, and Theory Constraint

What a theory *does* in social cognition is to enable or guide some cognitive process or function within that domain. The theory-theory claims that understanding others is enabled and governed by the theories of psychology which we build or innately possess, and how we apply them; the theory is thus a theory of human psychology.[27] For example, understanding the concept of sadness and understanding which conditions might make a person sad, coupled with the understanding that others are relevantly like ourselves, could be considered part of a theory of human psychology. But what is it about information and the connections between it that is theory-making? According to a pluralistic understanding, our folk-psychological interactions often include us acting in

[27] There may or may not be innate mechanisms that function according to that which is describable by theory—this is included merely to highlight the range of possibilities.

accordance with behavioural scripts (such as when we make small-talk or formulaically order meals at a restaurant) as well as attributing mental states—it is not clear in what sense the loosely socially governed behavioural guidance in scripts could play a part in constituting a theory about human *psychology*. As such, we need to be clear on what a theory *is* in this domain.

Below, I will make a case for theories being structures of information that are given their structure by the connections entailed by their 'special vocabulary', though the nature of the special vocabulary and the domain of the theory itself is guided by the information that makes up the content of the theory. A theory is thus both the information constituting the domain of the theory, and the connections between the information that gives it structure. We will see that there are various structures a theory might have, based on the kind of information it contains. As such, theories are multiply-realisable, despite possessing vastly different internal structures. The outcome of this is that when certain authors contend that their accounts of folk psychology are in-between theory and simulation, or are otherwise alternatives to them, on my understanding of theory they are giving accounts of different theoretical structures within the plurality of our folk-psychological practices; hence, this serves to clean up the dialectic. This is not a detriment to those theories; it merely means that theory and simulation are ultimately 'the only games in town' as the basic operations for socially cognitive processing.[28]

3.1 Properties of Theories

Theories appear to be formally structured; this gives us a starting point for the analysis of the core properties of theories. This is so even in examples of the weakest sense of theory: an informal notion in a conversation. Consider opinions given in response to questions: "I wonder why jeans are typically blue and not green?", "Well, my theory is that people like the colour blue more than the colour green." Such banality highlights the fact that, even informally, there is an implied structure to the relevant information. The information is that blue is more common as a jeans colour than green, with the theory explaining and contextualising this information by asserting that people just prefer blue over green. Another common use of theory, of course, is a scientific theory.

Considering theory as being formally structured will be useful in assigning some conditions for what ought to count as a theory in the domain of social cognition. In the

[28] À la Stephen Stich: "The theory-theory is not the only game in town, but it is the only other game in town" (Stich 1998, p. 145).

mind, the information in a theory is not necessarily represented in a natural language like a scientific theory would be, and the full contents of the theory itself may not be consciously accessible. However, I claim that we would not call such information a theory if it did not feature at least some properties shared with the scientific notion. To that end, below I will share Gopnik and Wellman's (1994, p. 259) useful summary of what they take to be core features of theories.[29] I will expand on why I think they are necessary, then add a fifth condition to flesh out the notion of the required 'coherent relations' between the objects in a theory.

3.2 Theory-making Conditions

1. Theories are always constructed with reference to evidence, some layer of experience different from the theories themselves.
2. Theories should lead to distinctive interpretations of evidence.
3. Theories produce interpretation of evidence, not simply descriptions and typologies of evidence.
4. Theories should involve appeal to abstract unobservable entities, with coherent relations among them.

This first point I take to be of little relevant interest, but it is necessary. In a theory of gravity, the claim that mass attracts mass is not useful for prediction or explanation unless such a thing could be observed pre-theoretically, thereby putting into action the test of the claims of the theory. Likewise, in a theory of mind, an understanding that 'crying people are feeling upset' is incoherent unless there are prior-experienced relations between emotions and behaviour.

The second condition notes that theories are able to explain events by principled distinctions from explanations of other events: the ball fell due to gravity, rather than ghosts. Without this ability, it would be impossible to make any kind of theoretical inference.

[29] My use of Gopnick and Wellman's work is not an endorsement of their 'scientific' understanding of folk psychology. For example, in the Stanford Encyclopaedia page for *folk psychology as a theory*, Ravenscroft (2016) splits discussion of 'varieties' of theory-theory into 'modular', 'scientific' and 'model'; the scientific variety is based on Gopnick and Wellman's work. We shall see that in using the 'scientific' variety as a springboard for understanding theory-making conditions, my understanding of theory is inclusive of all of these notions.

The third point is made with regard to Gopnik and Wellman's distinction of the organisation of experience into two different types—that of empirical typologies and generalisations, and that of "theories" (p. 260). Empirical typologies are cases, in their words, in which "orderings" and "glosses" of experience are couched in the same "basic vocabulary" as the evidence itself (ibid). As an example for basic vocabulary: 'trees with rough bark'/'trees with smooth bark', with an example of a generalisation being that 'plants try to face towards the sun'.

A basic vocabulary is not sufficient for theory-making because these descriptions and typologies do not produce an *interpretation* of evidence—these are just descriptions of the information that particular theories explain.[30] What makes a theory go beyond the objects of experience is (in part) an appeal to abstract, unobservable entities (with coherent relations among those entities)—this is the fourth condition.

This fourth condition I take to be crucial in order to understand what a theory is. Unlike empirical typologies and generalisations, theories create structures that explain generalisations, typologies, and observations by way of abstract entities and forces. To do this, they use a vocabulary that differs from the phenomenon itself—a *special vocabulary*. For example, a smooth-barked tree might be a *beech* tree because it has a distinctive genetic lineage that differentiates it from a *horse-chestnut* tree with its rough bark. Similarly, we might postulate *bacteria* as an unseen entity to explain the spread of an infection, or *special relativity* for certain interactions of light.

If we recall the discussion of the behavioural schemas associated with ordering in a restaurant, we can see how an organisation of this information would be theory-making—taken alone, these social rules are just empirical typologies and generalisations. With a special vocabulary that relates these rules to human psychology (such as understandings about desires not to be rude or commit a *faux pas*), the special vocabulary helps us to understand the interaction and participate in it.

To meet the four conditions is to be theory-making, but for this to be clear we need to flesh out the notion of the special vocabulary by defining the nature of the coherent relations between the content of the information given by the special vocabulary. Indeed, we want to know what sorts of relations and links between information the

[30] Whilst Gopnik and Wellman appear to take empirical typologies and generalisations as exhaustive of how the organisations of experience will be expressed, I am not committed to such a structure; we will see this in 3.2, on lawlike generalisations in theories.

special vocabulary is responsible for. Doing so will help us understand how folk-psychological accounts that posit themselves as alternatives to theory and simulation are instead merely theories. To that end, I add a fifth condition to theory-making. In order to do this, I borrow some familiar discussion from David Lewis (1972).

3.2.1 Fleshing out the Special Vocabulary: T-terms

In 1972, Lewis set out what he took to be the meaning of theoretical terms; in so doing, he described the structure of theories. His discussion of 'T-terms' will be useful in showing how relations between elements of a theory can be constructed, even when the theory introduces new terms specific to the theory. Suppose that T-terms are terms introduced by the theory, whereas O-terms are terms that were understood before the theory was proposed.[31] Lewis claimed that T-terms are names, though crucially they can be the names of "individuals, sets of attributes, species, states, functions, relations, magnitudes, phenomena or what have you" (Lewis 1972, p. 253). In the relevant sense, T-terms will be terms that name entities that are relevant to the explanation and contextualization of the bodies of information we possess.

Lewis presented us with a story that introduces T-terms, his point being that because of the way T-terms function, we understand what the terms mean without definition. He asked us to imagine a classic literary murder mystery. Suppose that "X, Y and Z conspired to murder Mr. Body…" (p. 250). His point here is that in describing such a case, X, Y and Z are theoretical terms referring to the names of people in the story. This is never communicated explicitly, but this is known to us because of the *role* that X, Y and Z play. That is, because of the functional role that X, Y and Z play in the story, we know that these new terms refer to people, to the murder suspects. To flesh this out more, we might say that the entities named by the T-terms occupy certain causal roles, such that "they stand in specified causal (and other) relations to entities named by O-terms, and to one another" (p. 253).[32] In a scientific theory, the term 'mass' will occupy a specified causal role in relation to entities named by O-terms such as 'particle' and to other T-terms like 'force'. A conception of T-terms as defined by their functional roles

[31] Some parallel to basic and special vocabularies can be noted here.

[32] The 'other relations' that T-Terms might hold to O-terms are mentioned presumably to note that causal relations are not the only relations these terms might hold, though the causal relations (and therefore their functional relations) are the relevant ones for the theory.

in the theory in relation to O-terms and other T-terms can be applied to theories of mind, too.

Lewis took this to relate to folk psychology in the sense that he took folk psychology to be a 'term-introducing scientific theory', with the terms introduced—the theoretical entities—being the common platitudes that we hold in relation to the mind. As such, "the names of mental states are the T-terms" (p. 256). In essence, then, states of belief and desire are T-terms that play a functional role in a theory of human psychology; we can understand this special vocabulary by the way in which beliefs and desires function in that theory.

As such, in specifying the causal roles of the occupants of T-terms, one generates a theory of their function. This point can be made explicit such that the special vocabulary not only describes the relations between the bits of information which make up the theory but also specifies the functional roles of the information in the theory—in essence, the special vocabulary determines why a particular piece of information is included in the theory. In a theory of human psychology, knowledge of social norms and scripts concerning ordering at restaurants serve in such a theory because they enable, amongst other things, a successful social interaction in which nobody is being rude, and one's desires for food are balanced with the waiter's desire to perform their role in providing service, et cetera. As such, detailing what makes a theory in this way explains how pluralistic elements of folk psychology can be included under the theory umbrella.

I claimed above that we needed a fifth theory-making condition to clarify the nature of the coherent relations between elements of theories, so on the basis of the preceding discussion, I write the fifth condition thus:

5. Theories specify the causal roles of the elements constituting the theory, specifying the causal relations between such elements by the use of the special vocabulary.

The causal roles specified by the special vocabulary function by showing how generalisations link to action. Thus, we can take condition 5 to be a causal elaboration of Gopnik and Wellman's 'coherent relations'.

Summing up, then, theories are organisations of information that demonstrate how the domain of the theory functions through the relationships of the information contained in the theory. A key takeaway is that information in a theory comes in two kinds: information that makes up the domain of the theory and information which makes connections between the information, generating a structure that essentially justifies why that information qualifies under the domain of the theory.

3.3 Constraining the Landscape of Socially Cognitive Accounts

I will now give some extra justification for my perspective on theory by touting its benefits in relation to how others have been using 'theory'. For example, Nichols and Stich's book (2003) outlining their theory-theory account claims that the relevant information that guides a (socially) cognitive process is merely "a theory-like body of information that serves as the basis for our understanding the mind" (Nichols and Stich 2003, p. 2). They also refer to such theories as "information bases" (ibid). On the view given by Nichols and Stich, the information and the relations between them appear to serve as inputs to cognitive mechanisms, the mechanisms of which may not even be wholly constitutive of a reasoning process.

I do not see much of a disagreement over the nature of theory here. In fact, Nichols and Stich seem to be answering the question of what it means to deploy a theory, as opposed to what theories themselves are. If an information base, plus relevant information from experience, is funnelled through some cognitive mechanism that may constitute part of a reasoning process, then this is just a more fine-grained approach to how theory might be realised in a reasoning process (whilst ironically being vaguer on what actually counts as theory). Charitably, we might suppose that the cognitive mechanisms that operate on certain inputs operate in a way describable by some theory—not to mention that this information needs to be the *relevant* information, and so there must be some principled way of selecting the input for the mechanism. As such, it seems that the notion of theory given by the five conditions above helps to guide the structure of what other authors take to be the information bases that make up the body of our folk-psychological knowledge.

Another benefit is that this presentation of theory helps draw the complex folk-psychological literature together into a more structured, pluralistic collection of views. This is because theory is not committed to a particular organisation of its structure; rather, it is merely asserted that one can form and specify some relations between its content and appeal to abstract entities and forces. For example, a theory that takes information and organises it into various models within its domain is a type of theory structure that differs from taking information and organising it according to the content and relationships between social narratives. In essence, the claim is that what a theory *is*, is in part to have a special vocabulary that dictates the relations between the information in the domain it covers, but it is neutral on how that structure might be realised in those that possess theories. To expand on this justification, I will briefly discuss an example of an alternative account of mindreading within the space of social cognition, in order to show that it nonetheless employs theoretical structures on my terms.

3.3.1 Theory in the Narrative Practice Hypothesis

Daniel Hutto's *narrative practice hypothesis* (Hutto 2008b, 2008a) takes folk-psychological understanding to emanate from experience of constructing and participating in narratives with those acting as teachers (or engaging in narratives whilst holding a pedagogical mindset). Take, for example, the lessons imparted in tales of Little Red Riding Hood. Hutto takes his narrative practice hypothesis to be an alternative to theory-theory insofar as theory-theory claims to be responsible for the core structural basis of adult folk-psychological competence.[33] His view of constructing narratives is meant to provide a better way of understanding how others act for reasons other than mere theoretical inference (Gallagher 2020, p. 166).

My claim is that exposure to folk-psychological narratives to "become familiar with the core structure of folk psychology" (p. 117) requires the storing of the relevant information and its relations acquired in those scenarios and utilising them later in their own narratives. Hutto claims that "Such stories familiarise us with the forms and norms of folk psychology" (Hutto 2007, p. 48)—those forms and norms need to be stored so that they can be applied later (otherwise nothing is learned), and as such they are represented in the mind. This is the construction and application of theory with a special vocabulary, in my view.

This is not to claim that the structure of folk psychology as a whole requires a special vocabulary related to narratives—rather, in Hutto's view, it is just that social pedagogy (enabled by participation in narratives) requires a special vocabulary related to narratives for "understanding them in a wider context; by acquiring the narrative that fills in or

[33] Hutto's target phenomenon is folk psychology as "predicting, explaining and explicating actions by appeal to reasons of the sort that minimally have belief/desire pairings at their core" (Hutto 2007, p. 44); my purview of folk psychology is broader, as it includes sub-personal processing in this domain and does not require reasons to be given in natural language.

fleshes out the particular details of that person's story" (Hutto 2007, p. 45). This is whether or not the subject is aware of this.[34] As such, whilst it is touted as an alternative to theory-theory (and its historical reliance on belief/desire reasoning), we can see that it is indeed just a pluralistic, theory-based, socially cognitive account.

To summarise section three of this chapter, I have explicated my understanding of theory, of theory-theory, and of the features of theories that are relevantly theory-making for my purposes; furthermore I added a useful constraint to the literature. As such, the literature can be carved into a dichotomy between applications of theory and applications of simulation, given that the attempts to construct accounts between and outside this space are still theories, in my view. This serves to streamline the discussion of the relevant accounts and constrains the resulting accounts of character reading to accounts that deal with theory and simulation. Of course, what simulation actually is has yet to be discussed; hence, it comprises the last section of this chapter. It should be noted that for the rest of the dissertation, my focus will mostly be on simulation's place in character reading, given the somewhat uncontroversial existence of theory in such accounts. In sum, I have given a detailed account of theory here, such that one can plug this understanding into the resulting hybrid account I endorse for character reading in Chapter four.

4. Simulation and the Threat of Collapse

Gordon (1995 p. 727), paraphrasing Hume, wrote that "our minds are mirrors to one another: they reflect one another's passions, sentiments, and opinions." The basic idea of simulation begins with our experiences. Often, it seems to us that an understanding of others can come from taking on others' perspectives, from seeing the emotions of others and recognising them for what they are because they reflect something in oneself. Crucially, the thought is that such mirroring could occur without the guiding hand of inferences informed by theories. This section outlines two types of simulation, and whilst both are aspects of simulation generally, it is the latter—'process'

[34] Hutto need not even disagree with my clarifications here—he notes of the empirical evidence for theory and simulation that "a careful review of existing experimental data at best shows that both simulative and theory-like generalizations come into play when we make third-person speculations about the contents of other minds, but noting this does absolutely nothing to settle the debate between TT and ST about what forms the structural basis of our FP-competence" (Hutto 2008b, p. 183). Hutto's complaint is not that theory is used, it is that he cannot see how theory (and simulation) alone give the structure for our acquiring folk-psychological competence. In my view, he can accept that simulation and theory come into play but couch the structural basis for folk-psychological competence in the structure of our narrative understanding given by its special vocabulary.

simulation—that I am concerned with in this dissertation; the former—'experienced' simulation—is a phenomenon that I take to be a fact that needs explaining, whether it be by theory or by process simulation. First, I present a quick primer on the common themes of simulation.

Common to all forms of simulation are themes of replication and resemblance. Simulation is the replication of some target, with the power and usefulness of the simulation being assessed through some necessary resemblance. Simulation achieves the same outcome as theory (outputs of practical reasoning, for example) by reproducing some crucial functioning for which a theory would be required to describe operations. For example, in reasoning about others' mental states, we might witness someone being berated and wonder how they feel. A theory-based answer might consult rules like 'nearly nobody enjoys being berated'. A simulative answer would replicate contexts and feelings to whatever extent we are able, thereby taking on others' perspectives to feel what it would be like to be berated, i.e., awful! Thus, our attribution of 'upset' is due to a sufficient resemblance between the target's context and the replication by the reasoner. Of course, simulation so vaguely defined is not precise enough. I mention replication and resemblance as a base for thinking about commonality between all simulation.[35] Now, I move to a discussion of what I take to be two main types of simulation.[36]

4.1 Experienced Simulation

Experienced simulation originates with Jane Heal (Heal 1996, 1998). Her point is that simulation as a theory about cognitive processing (as it is taken to be in mindreading debates) is just one account of how 'simulation' generally might be realised, given that simulation is a *practical reality* that warrants explanation. This is apparent in two senses. Firstly, in Heal's notion of 'co-cognition':

> It is an a priori truth that thinking about others' thoughts requires us, in usual and central cases, to think about the states of affairs which are the subject matter of those thoughts, i.e. to co-cognize with the person whose thoughts we seek to grasp (1998, p.484).

[35] For a discussion of what it means to bear resemblance, and the value of more 'concrete' similarities in the simulationalist project, see Fisher (2006). For my purposes, resemblance can be taken in an ordinary folk sense—smiling and laughing people do not particularly resemble grumpy characters, for example.

[36] These are not exhaustive. For example, the sense of simulation meant in 'computer simulation' is not discussed here. One will be able to infer from the next sub-section, though, that I take computer simulations to be complex combinations of theory with process simulations.

To co-cognise, then, is to think about the same subject matter as someone else, where doing so is to make similar inferences between involved concepts. To think about the same subject matter as others is to simulate their thinking, where such a resemblance makes this a simulative claim: "it resembles it in that it calls on the same grasp of the subject matter and results in thoughts which exhibit the same patterns and linkages" (ibid, p. 491).[37] The practical reality of simulation is also apparent in that certain reasoning experiences *feel* simulative, as opposed to feeling like rule-following. For example, mindreading achieved by taking another's perspective: 'They just dropped their ice cream. What are they feeling? What are they likely to do? I will pretend to have dropped my ice cream and see what feelings and desires follow.' This feels more like simulation than consulting an internal rulebook, such as we might when playing chess. Of course, whether the sub-personal mechanisms generating these simulative experiences are simulations themselves is an open question—I consider such an objection in 4.2.2.

As such, I bring co-cognition and the phenomenology of simulation under the banner of *experienced simulation*. Simulation exists both as a truth about our experiences of being able to think the same things and make the same inferences as others, and as a particular phenomenology of experience that differentiates it from theory-following. As such, a full account of what simulation is and its relation to social cognition needs to be inclusive of experienced simulation. In mindreading, it is clear that there is experienced simulation and so simulation is a crucial explanandum. However, it might be that what underlies our phenomenology of simulation is sub-personal theoretical inference. This sentiment parallels Spaulding's, when she noted that phenomenology is not strictly evidence for any particular 4E account of cognition (Spaulding 2015). As such, I move to a discussion of process simulation, which is an empirical project, and can thus be investigated as to whether it *does* underlie experienced simulations.

[37] It has been suggested to me that co-cognition may presuppose a conceptual-role account of conceptual structure (see Carey, 2011). This is fine for my purposes, but if atomistic accounts of conceptual structure were correct (see, for example, Kwong, 2007), then one might object that two people could think of the same concepts without making the same inferential 'jumps' to other concepts (hence the phenomenon of co-cognition is not self-evident). However, this seems to confuse the ontology of concept possession with the practice of inference making. The atomistic theory of conceptual structure claims that the possession of a concept is merely the possession of a certain mental representation, which is not constructed from any inferential relations between other concepts. This is in contrast to 'epistemic' theories which take concept possession to *be* holding certain beliefs and making certain inferences. However, regardless of the ontology of concept possession, it remains uncontroversial that we can reason and make inferences between concepts—to replicate those relations is to co-cognise.

4.2 Process Simulation

Replication and resemblance are tied to the *process* of simulation. If some attributions of mental states to others is achieved by simulation, then the cognitive processes one employs to make the attribution will be replications of some other processes. Here, I use Alvin Goldman's definition of process simulation, as it generalises well across various conceptions of simulation in the relevant literature:

> Process P simulates process P' = df.
>
> (1) P duplicates, replicates, or resembles P' in some significant respects (significant relative to the purposes or function of the task), and
>
> (2) in its (significant) duplication of P', P fulfills one of its purposes or functions (Goldman 2006, pp. 37-38).

Process simulation is an empirical project. One can use data from experiments to argue for the existence, or lack thereof, of simulative processes (both conscious and unconscious) underlying our practical reasoning about others' thoughts and behaviours. For example, if one mindreads by 'putting oneself in another's shoes', we might predict that certain failures in reasoning would arise from the imperfect replication of the target's processes. A good candidate here might exist in the failure to suppress one's own contextually relevant knowledge; one might look for evidence of simulation in mindreading by examining whether 'egocentric' errors are made in mindreading judgments. As a singular example amongst many (Goldman 2006, pp. 165–170), Hayashi and Nishikawa (2019) note that when subjects are exposed to stories in which a protagonist is either intentionally or unintentionally helped or hindered (depending on the task condition), they attribute greater levels of happiness/sadness to protagonists who were deliberately helped/hindered, regardless of whether the protagonist was aware of the intervention being deliberate. This purportedly demonstrates a failure to inhibit one's own knowledge of the situation in cases of emotional mindreading.

Process simulation in the domain of cognition sometimes focusses on the 'reuse' of cognitive mechanisms (Hurley 2008), where mechanisms are physical structures (usually in the brain) which implement multiply realizable algorithmic functions known as cognitive processes. Whilst the evidence for simulation in mindreading is contested, suppose that in visualising what a target sees, you might reuse your visual system for this purpose, rather than its usual 'seeing' purpose. The result of this simulation is a simulated mental state (of a visual experience). You might then use this simulation to mindread the target—for example, if you realise through visualisation that they cannot see your hands under the table, you can expect them not to have any particular knowledge about what you are holding. What might evidence this? Take it, for example, that whilst lobectomies of the occipital lobe (excising a particular part of the brain) affect the recipient's perception of their visual field, they also affect the size of their imagined visual field (Farah, Soso, and Dasheiff 1992). Furthermore, people who suffer from 'spatial neglect'—where they are unaware of objects on the left side of their visual field (often following strokes)—also tend to neglect the details of the left side of familiar scenes when describing them (Bisiach and Luzzatti 1978). Spivey et al. (2000) noted that "even when participants' eyes are closed, they tend to move their eyes in directions that accord with the directionality of the scene being described" (p. 5), suggesting that they are following the scene in their 'mind's eye'. Together, studies like these can be used to build arguments that simulations can occur in mindreading when cognitive processes reuse cognitive mechanisms to replicate someone else's cognitive processing. This is the sense of simulation to which I appeal in this dissertation, particularly regarding the neural reuse of memory systems, as detailed in Chapter four.

It is these notions of theory and simulation that we can use to evaluate the processes involved in character reading. Process simulation receives the most attention in accounting for the psychology of character reasoning in this dissertation, given the scepticism of its involvement (as detailed in Chapters three and four). However, it would be fruitful here to examine why exactly simulation is importantly distinct from theory. My classification of theory conglomerates many accounts under the theory banner; some might wonder why simulation ought to be treated differently. Indeed, such scepticism has not gone unexpressed in this area. Below, I will present Daniel Dennett's argument that simulation collapses into theory, then examine existing responses, defend simulation's distinctiveness, and finally show why simulation does not collapse into theory (even on my broad designation of theory).

4.3 Simulation into Theory: The Threat of Collapse

The threat of collapse was seemingly first articulated by Dennett (1987). He noted that if one were to imagine oneself as, to simulate being, a suspension bridge, any reasoning that follows (given that premise) depends entirely on one's knowledge of bridges and how they relate to physical laws like gravity.[38] Dennett claims that such knowledge needs to be organised into "something like a theory", and therefore that simulation

[38] If one has trouble *imagining* themselves as a bridge, one may merely *suppose* that they are a bridge.

collapses into theory. His point is that one's mental simulation of being a suspension bridge is constructed from not only those bare facts about bridges, but also how bridges react to the environment according to generalizable principles like gravity and tensile strength. This results in a 'folk psychology of bridges' with rules like 'bridges will not collapse because of wind but they might do because of hurricanes'; as with the bridge, so with simulation in mindreading, according to Dennett (pp. 100-101). The complaint is obvious as it pertains to my designation of theory. If simulation depends on a particular structure of information that is organised according to the broad theoretical conditions given, then there may not be much motivation for treating simulation as 'the other game in town'.

Goldman (1989, 2006) and Heal (1986, 1998) took issue with Dennett's objection. They claimed that pre-theoretical simulation could occur when the targets' states to be simulated are relevantly similar to the host system, such that no knowledge from theory is required to prescribe the conceptual and functional links in such processing; this similarity would apply for simulative mindreading. However, Jackson (1999) argued that the host's recognition of being relevantly similar requires holding beliefs (tacitly or otherwise) that constitute a theory about the 'similar' domain; thus, the problem persists.

In order to resolve the threat of collapse in my favour, I will examine the particularities of Goldman's most recent (2006) defence against collapse and defend what I take to be its strongest argument: the *feature, not a bug* argument.

4.3.1 Goldman on Collapse: Objections to a 'Resemblance to Self' Premise

Goldman makes several claims in his defence against collapse—I will discuss three of them, and endorse the third:

1. Whether you are justified or not in making a simulation (on the basis of lacking a belief about others being like yourself) is not relevant to the reality of simulation.
2. It is not clear that 'resemblance to self' is a relevant categorisation property that distinguishes agents from surrounding environments (in order for them to be targets of simulations).
3. A collapse into theory is not a total collapse—simulation still retains a distinctive character from theory.

Regarding claim 1, Goldman interprets the collapse critique as claiming that one would not be *justified* in making a simulation without the recognition of such a law of similarity between humans. However, Goldman notes that this is a different issue from how simulation actually works—it can certainly function (it turns out appropriately) without such a law, even if one's epistemic reasons for creating the simulation are weak. The problem with this response, however, is that one still needs some categorisation of the elements in one's environment in order to perform certain functions, such that the 'relevantly like me' belief may be required in order to categorise the agentive element of the environment as an appropriate target of simulation.

Indeed, Goldman seems to anticipate this response, as he goes on to make claim 2: of course, there must be some categorisation properties, but it is not clear that the relevant property (*a priori*) is 'resemblance to self'. The problem with claim 2 is that this is all it is—a claim. Whilst it is true that *a priori* it is not clear, very few things *are* obvious *a priori*. Indeed, a lack of being obvious is less convincing when we note that, in fact, we *do* have beliefs that others are like us. Claims 1 and 2 are not, therefore, convincing responses to the threat of collapse.

It is the slightly concessionary claim 3 in which I think a plausible defence lies. This is essentially the notion that simulation may require some theoretical knowledge and inference in order to function in human brains, but simulation is still clearly distinct from a theoretical inference itself. Goldman suggests that this is true along two dimensions, which I note here:

- "Even if the final stage of a simulation routine involves an inferential step utilizing a theoretical premise, this would not eliminate the distinctively simulational character of the earlier stages" (p. 30).
- "Simulation could be causally responsible for mindreading even if theorizing is also at work, because theorizing might simply implement simulation rather than replace it" (p. 34).

I take these quotes to describe two different occurrences of simulation with theory: the first describes simulation as *process* simulation, and the second describes simulation as *experienced* simulation. In the first, the idea is that in order for the relevant socially cognitive inference to be made, some sub-step of the process involving theory must occur in addition to the simulation. Such theories would be necessarily 'hybrid' theories. In the second, the claim is that the experience of some replication and resemblance would be the phenomenally felt output of sub-personal processes that are entirely theoretical. This, I suggest, is a claim that all interested parties would find plausible and is not problematic given the reality of experienced simulation.

Both of these claims amount to saying that theory being involved in simulation is a *feature, not a bug*—that is, the presence of theory is entirely within the expected bounds of the enaction of simulation. Claim 3 is actually plausible as a defence against the collapse of theory into simulation because it notes that there is only *partial* collapse—simulation retains its distinctive phenomenal character despite any theoretical underpinnings; it keeps its empirical designation of being a process that replicates and resembles some other process. Such a view essentially entails a necessarily hybrid socially cognitive account when simulation is involved, but this is no problem for me or Gordon, given that we endorse hybrid accounts.

4.3.2 Does Simulation Collapse into my Notion of Theory?

There is one final objection that might be levied against my attempt to defend against the threat of collapse. One might note that in deliberately giving a broad designation of theory (so as to simplify the categorisation of the literature), one might wonder why simulation counts as the 'other game in town' whilst something like 'narrative practice' is subsumed under my definition of collapse. Essentially, one might suggest that I have made the threat of collapse stronger.

My response is that, by my designation, simulation is not a theory because simulation is not information or connections between information; rather, it is a process (that manipulates information) that resembles and replicates some other process. As it is not the information or the connections between the information, it does not have a special vocabulary; hence, it is not a theory. Narrative practice, of course, does have these features. Perhaps one could object that, in describing a hybrid theory of social cognition, simulation itself features as an element of the special vocabulary that specifies certain roles of processes within the theory, but this serves to describe the hybrid theory, not simulation. it is notable that the role that simulation plays in the special vocabulary of a wider hybrid theory is distinct from the role that theoretical inference plays in such a theory—a difference which is not apparent for narrative practice.

Conclusion

In this first chapter, I have articulated the relevant technical terms used in this dissertation: social cognition, folk psychology, and mindreading. I justified these definitions beyond mere stipulation—I argued that my definitions were a better fit for the literature and presented fewer opportunities for confusing crosstalk, particularly in a way that will differentiate mindreading from character reading (as will be detailed in Chapter five). Secondly, I articulated key notions of theory and simulation in social cognition that will underpin my account of character reading: I first presented an account of theory that streamlines the discussion of theory and simulation in the literature; then, I articulated two relevant notions of simulation. Finally, I defended simulation against the threat of collapse into theory on the grounds of biting the bullet with a (crucially) *partial* collapse, but also noted that—given my wide definition of theory—simulation does not meet 'theory' criteria. As such, I have set up and defended the technical landscape for my account of character reading.

What is a Character Trait?

1. Introduction

Grumpy. Kind. Industrious. Lackadaisical. Generous. These are examples of traits of character, but we need to know why these words count, we need to know what a character trait *is,* if we are to understand the target phenomena in character reading. Take the following definitions of character traits from previously cited psychologists and philosophers:

- "Traits are stable personal qualities that moderate perception and action across a range of different situations" (Matthews 2015, p. 871).

- Traits are "dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and actions" (McCrae and Costa 2003, p. 25).

- "Personality traits are properties of a person that are taken to be stable, and they are often used to describe behavioral dispositions" (Andrews 2008, p. 16).

- "[C]haracter traits are temporally stable mental properties that relate to action in an opaque general manner across a wide range of situations" (Westra 2018, p. 1219).

- A character trait is a "relatively stable and long-term disposition to act in distinctive ways" (Harman 1999, p. 2).

Though this is a small sample of the definitions of traits in the literature, it is clear that these have much in common. That character traits are properties seems uncontentious, though it is debateable whether they are properties of humans alone, or also of some non-human animals.[39] Likewise, there seems to be some consistency or stability of these properties. This stability applies across situations and concerns dispositions towards certain behaviour (though, whilst I agree that character traits are described dispositionally, they are not themselves mere dispositions). What they do not all agree on, though, is whether character traits are also dispositions to perceive and experience in certain ways. I include this in my definition; hence, I will provide a defence of their

[39] See the introductory chapter, 2.3.

place. To begin, though, I present the definition of a character trait that I am working with.

A character trait is a temporally and situationally stable psychological ‘tendency’ of an intentional agent. Tendencies are composed of dispositional properties $X_{(1\ldots2\ldots n)}$, and an implicit history and/or summary of past $X_{(1\ldots2\ldots n)}$ manifestations. For a tendency to be of the psychological kind (and not merely behavioural) is for X to at least describe dispositions to perceive and experience, in addition to expressing behaviour, in certain ways across sets of situations.

Despite the technical terms, this definition is constructed from the ways in which character traits are usually conceived of in various canons, though with a couple of marked differences: that character traits as tendencies are not mere dispositions, and that certain ‘inner phenomena’ are included in the concept.

This chapter details my account of this definition, in particular in terms of traits as tendencies. Section 2 details some relevant metaphysics for character traits—on their relations to properties and dispositions, and then raises and answers an objection of the causal efficacy of dispositions. Section 3 makes my case for character traits being more than dispositions, endorsing an account of traits as ‘tendencies’. Section 4 returns to the definition of traits posed above and fleshes out some of the remaining terms that require context, particularly the psychological dimension of traits.

2. The Metaphysics of Traits

Character traits are properties of individual people. We have seen that character traits are often described in dispositional terms by personality psychologists and philosophers. Characterisation in terms of dispositions is also a folk understanding of character traits. Whilst I hold that character traits are not dispositions; it is uncontentious that they are generally described dispositionally. For example, the character trait of courageousness can be described with a cluster of dispositions to act/think/perceive in certain (courageous) ways. To understand what a character trait is, we therefore need to know about dispositions and, crucially, the causal roles they play in producing behaviour. This is because predictions and explanations of events, by appeal to character traits, reference the causal roles that things play in bringing about the event. However, in order to understand the relation between dispositions and causal roles, we need to establish that properties can have causal roles.

2.1 Properties and Causal Roles

Somewhat counter-intuitively, what properties are exactly does not matter so much as what they can do. I claim that properties of things are factors in causation, such that character traits (as properties) will have causal roles in disposition manifestation. In essence, the character trait of courageousness plays a causal role in your manifesting courageous perceptions/thoughts/actions—who you are helps cause you to be a certain way.

The causal role of properties can be seen through an example from Mumford and Anjum (2011): Suppose we place a bag of fruit on a mechanical set of scales. The needle that measures weight will move accordingly. What caused the movement of the needle? The mere occurrence of the fruit on the scale? The event of placing the fruit on the scale? Something about the fruit itself? All three are involved in causation, clearly, but it is only the properties that the fruit has (its mass, in this instance) which enable the occurrence of the fruit on the scales or the event of placing them there to have anything to do with the needle moving. Of course, one might say this of any of the above causal factors—for example, without the mere occurrence of the fruit on the scale, the needle would not move—so we need another example. Suppose we have two identical glasses of water, and into each glass we pour a different powder. In the first case, the powder dissolves. In the second, it does not. Here, the *presence* of the powders in the water is the same, as is the *event* of placing the powder in the water. *Ceteris paribus*, it is the powder's properties (its solubility in water) that must play a causal role in the powder dissolving; hence, properties are not causally impotent.

I do not have the space to commit to a particular metaphysics of properties here, but since I reference Mumford and Anjum (2011) on causes, and because I also refer to Mumford (2003) on dispositions later, I want to clarify my position on properties relative to Mumford and Anjum's. They hold that "properties are just clusters of causal powers" (2011, p. 4). What powers are is contentious, and they intentionally do not offer a definition (p. 5), but it seems in general that they are using powers to be some kind of potentiality: causal powers are potentialities to cause effects; *they are dispositional*.[40] This appears to be confirmed when they note that "we are using the terms 'power' and 'disposition' as equivalent" (p. 5). As such, for Mumford and Anjum, the property of

[40] Potentiality is an Aristotelian notion. See Witt (2003) for more on Aristotle's notions of potentiality. See Vetter (2015) for an in-depth account of potentialities (and dispositions as kinds of potentialities).

'hotness' just is a cluster of powers to heat, to vibrate molecules, et cetera.[41] Therefore, if one's personal preference is for a powers ontology, this will work on my account too, though I need not be wedded to it—what is important is the reliance on properties being involved in causation.

In summary, returning to the relevance to character traits: character traits are properties, properties have causal roles, and those properties can be described dispositionally. The character trait property 'generous', for example, can be described as the disposition to act generously, such as tipping well in restaurants.[42] What, then, are dispositions?

2.2 Dispositions

I will list four general features that authors often consider paradigmatic dispositions to have, compiled by Alvarez (2017). This is so that we have a general idea of what a disposition is. I will return to these features in section 2, as I hold that traits are not mere dispositions: they do not meet all the paradigmatic criteria. I claim that traits lack the 'independence' feature of dispositions. Assume, for now, that traits are something other than mere dispositions.

The features of dispositions are as follows:

1. They are *directed*. A disposition is directed towards some outcome. For example 'fragility' is a disposition towards easy breakage; the disposition is defined by its manifestation.
2. They are *independent*. Something can have a disposition without ever manifesting it. A fragile object has the disposition to break easily. It has the power to do so, but it need not ever manifest. A fragile glass is still fragile regardless of its never breaking.
3. They have *stimulus conditions*. Dispositions manifest under certain conditions. Without the presence of the stimulus conditions, the disposition will not manifest. However, the mere presence of stimulus conditions is not necessarily sufficient for manifestation. For example, dropping a fragile glass is usually a stimulus condition for the glass manifesting its disposition to

[41] However, this is not the case for all properties. They are not talking about 'relational properties' or 'abstract properties', like being a kilometre away from the nearest gym (the former) or being divisible by two (the latter). They do not consider that these properties have causal powers.
[42] This ignores, for now, that my dispositional description of a character trait has an added summative feature of past disposition manifestations.

break easily. However, if it is protectively packaged in a box, it may not break when dropped.

4. They have a *categorical basis*. This is a property(s) distinct from the disposition itself that grounds the disposition—the property by virtue of which something has the disposition. For example, a glass may be fragile because glass is made of silicate molecules that are weakly bonded in a suboptimal pattern, which creates various microscopic cracks in the glass's surface such that the application of force widens the cracks and causes easy breakages.

One might consider dispositions themselves to be properties, or to be, as Hugh Mellor argued, features of predicates. If they are properties, then there is a worry that dispositions do not seem like the sorts of things that can play causal roles. If dispositionality exists at the level of predicates, rather than properties, then this issue will not arise. In the interests of theoretical neutrality, I will defend the properties interpretation against the causal roles worry, such that either option for one's preferred stance on the ontology of dispositions is compatible with my notion of a character trait. In my explanation of the predicate conception, I will note that the account implies that character traits cannot be mere dispositions—a claim I endorse and give arguments for in section 2 anyway.

2.2.1 The Causal Efficacy of Dispositions

Here, I outline and respond to a classic objection about the supposed causal efficacy of dispositional properties, which is brought to bear by the dispositional characteristics of traits. If dispositions are properties, and properties have causal roles, it follows that dispositions have causal roles. How can this be the case? [43] In particular, we might think that giving a causal explanation using dispositions is particularly unsatisfying. It is common in the literature surrounding dispositions to refer to *La Malade Imaginaire*, the 1673 comédie-ballet by Molière, in which Molière mocks those who explain, for example, how opium puts people to sleep by referring to its *virtus dormitiva* (dormitive virtue). This example is used to challenge how it can be that a disposition causally explains anything at all, when the reason that opium puts people to sleep is that is has a sleep-inducing power—that it has the *disposition* to put you to sleep. Similarly, to say that

[43] One might wonder why this is an issue if I do not claim that traits are wholly dispositions. The answer is simply that the part of traits that is not disposition is a part that is unsuitable for doing the relevant causal work, given as they are causal *histories;* hence, they are the wrong direction for causation to be relevant.

sugar is water-soluble is simply to say that it dissolves in water; an explanation of why sugar dissolves in water by reference to a disposition seems unsatisfactory. Relevantly, if dispositions are causally efficacious, a character trait would provide a cause for some trait expression; as such, that our explanations of such behaviour are genuinely explanatory might be in jeopardy.[44]

The dormitive virtue is a kind of triviality objection that poses a problem because it acknowledges, unlike *simple* triviality objections, that there is something *in particular* about, for example, sugar that makes it dissolve in water. In this sense it goes beyond a simple triviality objection that would claim that dispositional explanations are not explanatory at all, to which one can respond by distinguishing the particular cause from chance, and by noting that something about sugar—rather than some other substance—is responsible for the dissolving. My response to the dormitive virtue objection is, like Stephen Mumford, to bite the bullet on triviality as an explanation, but still insist on causal efficacy.

Firstly, on the triviality of explanation: we can concede that explaining what it is about sugar that makes it dissolve in water, or what it is about opium that makes one drowsy (as solubility and a dormitive virtue, respectively), are not good explanations. Similarly, for character traits, we might think that the answer "Because he is evil" to the question "Why did he kill those people?" is not particularly satisfying as an explanation on its own without any other information or context; dispositional explanations "provide no detail of the mechanisms involved in a disposition manifestation (…) [T]he property is characterised only functionally, not 'structurally'" (Mumford 2003, p. 26).[45] Perhaps the categorical basis of the disposition would count as a good explanation (recall the categorical basis of a glass's fragility), but certainly not the disposition itself. Without extra information, say about motives or what it means to be evil, this explanation is essentially trivial: "Why did he kill those people? Because that's what people like him do." I take it that this is not a particularly controversial bullet to bite.

What about causality? We can accept that such an ascription is uninformative beyond the simple sense, but that does not mean it is causally impotent. Stating that 'the cause of dissolving caused dissolving' *is* trivially true and uninformative, but we would not then reject that the cause of the dissolving caused the dissolving. Note that there is no

[44] I say 'a' cause, I am not committed to it being *the* cause.

[45] I give a structural account of this across Chapters four and five.

positive argument for dispositions as causes here. This is implicit in the understanding that properties have causal powers, and dispositions are properties. The *virtus dormitiva* objection only raises an objection of the triviality of *explanation* to motivate intuitions about the triviality of *cause.* It is in the poor explanation that a dispositional explanation gives which makes us ponder the causal efficacy of dispositions. However, a couple of examples in which dispositional explanations can be non-trivial have been given (in response to simple triviality), and Mumford demonstrates in the *virtus dormitiva* case that we can separate the triviality of explanation from that of cause; hence, our intuitions are askew. A particular way in which this bears fruit later is that in Chapter five I discuss trait explanations, where (crucially) some trait explanations may appear trivial. This does not mean, for the reasons above, that such explanations of behaviour imply that behaviours are not at least in part caused by one's traits and are therefore inaccurate.

2.2.2 Dispositions and Predicates

If, instead, dispositions were features of predicates rather than properties, what does this mean? For fragility to be a disposition, for example, is for the predicate 'is fragile' simply to express some set of conditionals that together are necessary and/or sufficient for the correct application of the predicate: X is fragile because *if* one were to drop it, it would break. Naturally, this is too simplistic a reduction—we can call more appropriate formulations the 'reduction' sentences of such conditionals (Mellor 2000, p. 761). Hence, the properties that X has are simply properties, whilst the dispositional feature exists at the level of the predicates, not the properties.

The case for dispositionality being a feature of predicates is made by Mellor (2000), who remarks that if this is true, then there is no reason to think that the corresponding properties of dispositional predicates are "all of a kind, and different in this respect from properties corresponding to non-dispositional predicates" (ibid, pp. 767-768). The main draw of this account, therefore, is that it nullifies any tension as to metaphysical differences between dispositional and non-dispositional properties. Hence, many different properties satisfy the predicate *is fragile,* such that all the predicate picks out are those properties that satisfy its reduction sentence.

It should be noted that not requiring commitment to some complex explanation of differences in kind between dispositional and categorical properties is not a worry that affects Mumford's property metaphysics. This is because on Mumford's view, almost all properties are dispositional in the sense that they are powers; thus, there is no particular difference in kind between categorical and dispositional properties—properties are individuated by their causal role, and not by whether they are categorical or dispositional.

It seems that an implication of this predicate view of dispositions is that character traits cannot be mere dispositions precisely *because* traits are properties (of people). Whilst this would certainly be a boon for my account, I give independent reasons for thinking that traits are more than mere dispositions in the next section, so I still have no quarrel with any particular view on the ontology of dispositions.

In summary, then, I have detailed and shown how none of the relevant idiosyncrasies of various metaphysics (of properties and dispositions) will negatively impact my account of character traits, and I have set the stage for the main claim of my account—that traits are more than dispositions.

3. The Tendency Account of Character Traits

3.1 Traits Lack Independence

We understand now how character traits are causally efficacious properties that can be described dispositionally, so this section explains how it is that character traits are not mere dispositions.[46] This is because whilst one of the paradigmatic features of dispositions (section 1.2) is that they have an independence feature, so what it is to be a character trait is to be characterised by *dependence* instead. A quick reminder of *in*dependence: this "consists in the fact that powers [dispositions] are ontologically independent of their manifestations: an object can have a power that is not being manifested, has never been manifested and will never be manifested" (Alvarez 2017, p. 72).

By contrast, Alvarez notes that character traits "are dispositions whose possession requires (ontologically) that the object display the sort of behaviour (broadly understood) that is characteristic of the disposition" (p. 79). Indeed, it seems to me that the notion that character traits must have manifested in order to be possessed is apparent in the way in which we use the term 'character trait' in real life. As such, my defence of character traits as 'manifestation dependent' consists in suggesting plausible

[46] Whilst traits are not mere dispositions, they *include* dispositions; hence, to make a claim like saying that generous people are disposed to X behaviour is simply to reference the dispositional properties (or reduction sentences) that in part compose particular traits.

grounding for dependence and explaining away objections to dependence. Take someone who has never once been brave in their entire life, not even thinking about acting bravely. It seems highly unlikely that folk would say that they might nonetheless be a brave person. This is also reflected in the literature; for example, Hacker (2007) notes that "Physiological dispositions of health apart, human dispositions are dispositions of temperament and of character. These are very unlike inanimate dispositions (…) a person cannot have a disposition of temperament or character and never exhibit it" (p. 119). I will respond to challenges to this intuition shortly; for now, I am concerned with detailing how traits might be more than mere dispositions.

3.1.1 Hampshire's Account of Traits

The earliest discussion of the dependence feature that I could find is in Hampshire (1953): he claimed that to make a statement about someone's character is to provide a summary of their past behaviour and tendencies to behaviour, and further, that this constitutes an important difference between a trait like irascibility and a property like fragility. Such a summary could not be provided if the disposition need not have ever manifested. This extra summary therefore constitutes the addition to the dispositional descriptions that I claim makes a trait not strictly a disposition. It is in this extra summative detail that the term 'tendency' will arise to replace 'disposition'.

There are two clarifications to make here, however. Firstly, Hampshire's terminology is somewhat opposed to mine. Hampshire took 'dispositions' to refer to traits, whereas he thought that properties like fragility ought to be considered as mere descriptions of the causal properties of things (Hampshire 1953, p. 7). However, I reject Hampshire's notion of a disposition, firstly on the basis that traits (as properties) do in fact have causal roles: a trait attribution is, at least in part, a description of the causal properties of people. Secondly, he falsely rejects the claim that dispositions are hypotheticals, preferring dispositions of character to be conceived of as categorical. This is because he notes that in calling someone generous, he does not mean that under certain specific conditions they would respond to the situation in certain specific ways (p. 10); the accuracy conditions of traits are different to properties like fragility (p. 8). However, whilst one might not *mean* this in making a statement, this does not mean that in possessing a particular trait, there are *no* manifestation conditions for the disposition; neither does it mean that disguised hypotheticals are not hypotheticals at all. Just because one does not know or intend to imply a particular reduction sentence for a trait does not mean that there is not one.

Indeed, traits do seem to have manifestation conditions, for which I give two reasons for thinking. Firstly, an example: suppose someone is a kind person. They find themselves in the familiar thought experiment in ethics—that which is usually used as an argument for the existence of objective morality: a child is in pain, and they can press a button to administer instantly effective painkillers to the child at no cost to themselves. Should they press the button? As this is a thought experiment, one can rest assured that all conditions are accounted for such that the only relevant consideration is whether to freely press the button at no cost. It seems to me that if someone genuinely is a kind person, they *will* press the button in this thought experiment. It may be that pressing the button is *also* a manifestation of their goodness as a person, their trait of mercifulness, or their code of ethics, but certainly the opposite must be false: a kind person would never *not* press the button; hence, one may count this scenario as at least one manifestation condition for the trait of kindness.

The second reason for traits having manifestation conditions relies on the premise of the physical determination of the universe—on an assumption of hard determinism, one can establish definite causal connections, even if those are part of a complex web of causality. Hence, if the universe is physically determined, and traits have causal roles, then they must have manifestation conditions, where those conditions must be determinable so as to fit within a causal history of the universe in which every effect can be traced back to a cause (and so on) until the Big Bang. Regardless, one can reject hard determinism and still agree with the former example (given that traits have causal roles); as such, I think we have multiple good reasons for rejecting Hampshire's specific terminology on that score.

The second clarification concerns his paradigmatic features of dispositions. Hampshire discusses the difference between *making a statement* about a disposition ('X is jealous') and *statements* about paradigmatic dispositions, which contain disguised hypotheticals (or in Mellor's language, reduction sentences). As such, much of his discussion is about those features that statements about dispositions should have and we might therefore wonder why we are licensed to talk of the ontology of traits in this manner, instead of merely to talk of statements about them. However, some of his conditions for statements that refer to dispositions contain explicit references to what he seemed to think dispositions actually *are*. For example, “A disposition must be manifested and must show itself in actual incidents; there must be at least some cases or instances of it dispersed over some period of time” (p. 6). This seems to refer directly to the disposition, rather than solely to a feature of a statement about dispositions, (though, notably, he rejects the dependence feature regarding paradigmatic dispositions). As such, while a statement about a trait is “a statement which summarises what tends to happen or is liable on the whole to happen” (p. 5), and a statement about a paradigmatic disposition is just a description of its causal history (for Hampshire), there is clear meaning in his writing that the ontologies of the two types of phenomena (traits vs. dispositions) differ; it is the *statements* about them both that *highlight* this; a summary or history of trait manifestation relies necessarily on a positive history of trait manifestation. It is this feature of traits that is distinctive from features of paradigmatic dispositions like fragility.

3.1.2 Objections to Manifestation Dependence

However, what about objections to traits having this dependence feature? Alvarez (2017) deals with a few such suggestions—I will present one response here and show how it succeeds, but the main response to objections of my account is forthcoming in response to the notion of tendencies themselves in 3.2.2. For manifestation dependence, some might object on the grounds that it is common parlance to say that we have ‘discovered’ a trait within ourselves, perhaps after some particularly noteworthy behaviour, such as courage in running into a burning building to save people. The use of ‘discover’ would suggest that we possess the character trait before the trait expression; hence, we can have a trait without it manifesting. However, this is not enough to claim that character traits need not be expressed in order to be possessed. Alvarez responds that perhaps the ‘discovery’ is just that the incident makes you see the point of courage, or, for example, some treachery sets you off on the path of becoming a treacherous person (p. 83). I think this is an intuitive way to conceive of it because people tend to accept that character traits are somewhat malleable over time and experience, despite their general stability. It would be weird to say that someone who has spontaneously engaged in treachery has been treacherous all along, despite prior behaviour to the contrary.

Of course, one might counter propose that people who have many treacherous patterns of thought (in planning their treachery) but do not engage in overtly treacherous behaviour until the opportune moment might nonetheless be thought to be treacherous. However, using the word 'overt' here seems to imply that some degree of implicit or covert treachery is fine. This accepts too much, as even hidden treacherous behaviour is nonetheless behaviour, and thought patterns of treachery are still inward expressions.[47]

Having noted how character traits cannot abide by the independence feature of paradigmatic dispositions, and that they in fact seem to abide by a dependence feature that requires a history of disposition manifestation, I now turn to defending my terminology of traits as *tendencies*.

3.2 Tendencies

Here, I directly apply Hampshire's notion of a necessary summary to character traits themselves, and not mere statements about them. Whilst it is true that statements about traits are implicitly summative, this merely demonstrates the correct application of language that involves traits. Of course, however, we do need to have a word to describe the conjunction of traits and implicit summaries. My claim is that 'tendency' is apt. For this, I follow Maria Alvarez in both definition and naming convention: "attributing a character trait is partly a record of past and present behaviour, broadly understood, but it also provides grounds (albeit defeasible ones) for predictions of future behaviour" (Alvarez 2017, p. 85). I should note that Alvarez gives this definition and calls it a tendency, but she does not seem too enamoured with the idea (ibid). I propose that we *should* adopt it, as it fits more neatly into both language and our common conceptions than she seems to think.

3.2.1 Tendencies in Ordinary Language

Alvarez's reluctance to adopt the term 'tendency' for traits is due to the fact that

> we should remember that the decision to call character traits 'tendencies' rather than 'dispositions', though reflecting a real difference between them and 'paradigmatic dispositions', would to some extent be a terminological choice that introduces a degree of regimentation relative to our ordinary use of these words (Alvarez 2017, p. 85).

I claim that this *does not* introduce the regimentation that she thinks it does on two grounds: firstly, it fits existing scholarship on the issue, such as that of authors who demonstrably use the term 'tendency' in this manner. Secondly, even for folk usage, our intuitions about the differences between the terms disposition and tendency are

[47] Justifications of inward expressions constituting trait manifestations are given in 4.2.1.

intelligible (though admittedly, Hampshire is an outlier). As to the first point, below is a small sample of authors using tendencies in the manner I suggest:

- "We thus have terms for dispositions that reliably and frequently manifest (tendency)" (Mumford and Anjum 2011, p. 5).
- "Something is prone to V or has a tendency to V only if it V's with some regularity—the concepts of a proneness and a tendency being frequency concepts." (Hacker 2007, p. 97). Hacker also says more on tendencies: "Tendencies and pronenesses are not, as such, human dispositions" (p. 119); "A fragile object may be liable to break if dropped, but this does not imply that it has a tendency to break if dropped (although things of that kind do have such a tendency)" (p. 118); "To describe someone as irascible, jealous or compassionate by nature is to describe his affective pronenesses and tendencies." (p. 115).
- The human tendency toward a confirmation bias leads us to look for reasons why a hypothesis is true, rather than for reasons why it might be false" (Andrews, 2012a, p. 124). This is an example of 'tendency' being used in context when not discussing traits. Here, the human tendency towards confirmation bias can be taken to mean that humans have a disposition towards confirmation bias, in addition to a history of past manifestations of such bias; hence, the claim can be made that it applies to humans generally. Without the disposition plus an implicit history, the talk of tendencies does not make sense here, given that the discussion is about a not-uncommonly manifested phenomenon.

Regarding the second point about ordinary usage in conversation, consider the following example. I might let my friend borrow my car for a trip, but I say to her: "Be careful, because of the way it has been built, the clutch has a tendency to stick". Consider now that I clarify: "I mean, it never *has* stuck, but it has the tendency to." This seems to be a strange clarification, whereas had I said "Be careful, because of the way it has been built, the clutch has a disposition to stick—it has not stuck yet with me, but you should be careful", this, I claim, would not elicit the same incongruity in an ordinary understanding of these terms.[48] As such, I am not sure what regimentation

[48] It is entirely possible that in thinking about this a lot, my own intuitions are skewed. That said, I have tested this example on both native and non-native English-speaking folk, and they reported the same. These were obviously small samples, so perhaps there is an opportunity here for some experimental philosophy to be done.

Alvarez is referring to when it comes to tendencies, as it seems to me that the term is already in usage in the manner described. There are further objections we might want to entertain regarding traits as tendencies though, which are the focus of this next subsection.

3.2.2 Objections to Character Traits as Tendencies

There are three objections to character traits specifically as tendencies, which I will respond to here. The first objection can be phrased thus: "Does this mean that I can have a courageous disposition without having the character trait of courageousness (if there is no implicit summary given)? That does not seem right." What this objection seems to be pointing to is that the necessary addition of an implicit summary might reduce the definition to absurdity in suggesting that someone can have a disposition to be courageous and yet not have the courageous trait.

To this, the response is to say that it is not clear that the sense of disposition in use here actually refers to paradigmatic dispositions as I have been using the term. Disposition is often used as *shorthand* for temperament—one might have a 'sunny disposition', which obviously does not mean that one has a disposition to be sunny.[49] Such an attribution means that they have a temperament that is characteristically upbeat, where such a temperament is noted after it has, crucially, manifested.

As such, this objection is not an objection about the aptness of traits as tendencies so much as it is a potential counterexample to the ordinary language distinctions between dispositions and tendencies that I claimed in 3.2.1. Therefore, this example seems to be a case in which one, in ordinary language, uses the term 'disposition' essentially to mean 'tendency'. However, I think this can be resolved by noting that distinctions in ordinary language-use for particular terms are important for demonstrating the value of using different terms, but we should not rely on particular examples of how people *can* use words as evidence against *conventional* differences between certain terms. In essence, the fact that there is a demonstrable difference between usage for talk of properties of dispositions and tendencies, but not for colloquial 'ways of being' such as having a sunny disposition, is acceptable. This is due to the fact that the former discusses the details of what makes up traits and therefore the difference in word usage needs to reflect a real distinction being made; in the latter case it is unproblematic that people

[49] Something being 'shorthand' for something else, and thus meaning something else, is an important phenomenon that garners attention in Chapter five.

might not know (or care about) a technical distinction between dispositions and tendencies. In essence, we should not expect such language-use to always conform automatically. Indeed, if they needed to, then I would not have even been *able* to use the car example above as expressing a 'weird' clarification about tendencies.

One may respond that fine, maybe we cannot have a courageous disposition without possessing the trait, but a second objection one could levy against traits as tendencies is that my account permits that one (in theory) might have a disposition to engage in actions that are *characteristic of* how courageous people act but not have the courageous trait. A notion of a trait that allows this does not seem fit for purpose. This is on the grounds that general dispositions towards actions that are characteristic of a trait are essentially dispositions to act as if one has that trait; to deny possession of the trait at such a point seems an absurd conclusion. To this objection, I respond that there is a meaningful difference between, for example, courage the *behaviour* and courage the *trait*. Similarly, there is a difference between a happy emotion and a happy trait—the former is fleeting and the latter more enduring. As such, this distinction is still warranted, as it is perfectly plausible that people *qua* people are disposed to actions that are characteristic of how courageous people act, as humans are disposed to a whole host of behaviours that are describable under different categories.

The final objection against traits as tendencies that I will discuss is to claim that because a summary of past behaviour might be just one instance, we can therefore presumably possess a trait if we have only expressed the behaviour once. This seems too quick, given the supposed enduring nature of traits. I think, however, that this is an easy bullet to bite. Yes, one might be said to possess a trait despite one manifestation, but this is not surprising. Context is important: we can imagine cases in which this one manifestation of behaviour is particularly impressive or extreme in its manifestation, where in its enaction we might even expect further manifestations of that kind. Take someone who has been bullied unrelentingly for months. One day, they publicly stand up to the bully despite the bully being backed up by their cronies. It takes courage to do something like this: because the odds are so stacked against our victim here, they become courageous in this manifestation of behaviour. Indeed, we might now expect further instances of this kind—if they had the courage to stand up to the bullies once, despite being outnumbered, we would not naturally expect them to never do anything of the like again. This is not to say that they *must* do something alike again. If they continue to be bullied and never stand up for themselves ever again, we might reassess our trait attribution, but the point is that one instance is potentially sufficient to be said to possess a trait.

In summary, character traits are usefully and correctly described as tendencies, which are dispositions plus an implicit history of trait manifestations. Mere dispositions need not manifest, such as fragile glasses that need not ever break, but one must exhibit the characteristics of a particular character in order to be said to possess the trait. I took Maria Alvarez's suggestion that traits are usefully distinguished as tendencies and showed that it was a distinction based on more than just a terminological choice; then, I defended the idea of traits as tendencies from potential counterexamples. Traits as tendencies are the key part of my account of what a character trait is and have thus received the most attention here. However, the definition at the start of the chapter included a little more than this, so for the final section I return to this definition to discuss the justification for its other parts, particularly the relationship between traits and psychological states.

4. Defining Character Traits

Firstly, a reminder of the definition of a character trait that I am working with: A character trait is a temporally and situationally stable psychological 'tendency' of an intentional agent. Tendencies are composed of dispositional properties $X_{(1\ldots2\ldots n)}$, and an implicit history and/or summary of past $X_{(1\ldots2\ldots n)}$ manifestations. For a tendency to be of the psychological kind (and not merely behavioural) is X to at least describe dispositions to perceive and experience, in addition to expressing behaviour, in certain ways across sets of situations.

I will take it for granted that I need not explain, for example, what an intentional agent is (I mean any reasonable commonly understood notion of it). The terms that I think still warrant explanation are traits being 'stable' and traits being tendencies that are psychological (in addition to being behavioural).

4.1 Traits as Stable

It is generally held that character traits are in some way stable. What is meant by this is that they are stable temporally and situationally. A generous person today is likely to be a generous person tomorrow (or even in many months). The tendency to manifest one's traits is also generally stable across situations (though caveats below).

However, it is perfectly possible (and in fact common) to act in trait-inconsistent ways, and traits can indeed change over time: extroversion decreases with age (Borkenau and Tandler 2015, p. 920), as does openness and conscientiousness (Lucas and Donnellan 2011, pp. 853–854). The stability of traits merely implies the probability of displaying trait-consistent effects on behaviour. Whilst we may be tempted to describe stability as a disposition to manifest with a certain frequency, note that the implication of stability is one of dependence—the notion relies on past disposition manifestations, and as such forms part of the implicit summary of a tendency. Perhaps tendencies might turn out to be stable by definition, but as it stands there is nothing important that turns on whether we call a character trait a 'stable tendency' or a 'tendency' that by definition might include stability.

What might these complexities of trait expression mean for a robust concept of traits as 'stable'? Indeed, it is too much to say that making trait-based predictions on a situation-by-situation basis is going to be fruitful. Both Andrews (2008) and Westra (2018) note that predictions of behaviour on the basis of character traits are not particularly accurate. As such, whilst (in a general sense) traits are stable across situations, it is a more accurate definition to generalise across 'sets of' situations; your character trait predictions of a work colleague may be accurate across work situations but decidedly less so in leisure settings. Despite this inaccuracy of predicting behaviour across differing situations relative to some others, as Andrews notes, "Because people understand traits as stable and constant dispositions of an individual, they are used to make predictions of behavior" (Andrews 2008, pp. 16-17). As such, when I talk of stability here, I am referring to a folk conception of traits as stable.

I do not claim that the folk are misunderstood about traits being stable, for empirical reasons cited below. I employ a (vague) folk notion of stability here in order for my work to truly track the relevant phenomena of 'character traits' that people are familiar with, as opposed to some specialist philosophical artefact. However, doing so invites both an objection about how we could make trait predictions on a pluralist account, and a worry about the meaningfulness of a notion of stability, detailed below.

In Westra's 2018 paper, he challenges Andrews' pluralist notion of character traits. This is on the basis that if the pluralist notion, which does not include any mindreading in how it is employed in social cognition, is to represent situations for the purposes of trait predictions, then a dilemma may be raised for how these traits are parsed. The dilemma is this: If traits are parsed with a broad representation of the situation, then our trait predictions would be even more inaccurate than they seem to be, for we would predict generous people to, for example, tip in restaurants even after especially poor service. On the other horn of the dilemma, where situations are represented narrowly, then even mostly familiar situations will be treated as novel; hence, existing trait-behaviour associations could play no role in making the prediction (Westra 2018, p. 1225). As such, the pluralist has a problem in accounting for how trait predictions can actually function.

The reason that this is a problem for me is that in employing a folk notion of stability, I cannot escape the dilemma by suggesting some middle ground between a broad and narrow representation of the situation, precisely because I am employing a vague notion of stability that corresponds roughly to what the folk would endorse.[50] Two questions arise, then: with a basic notion of stability, how can I endorse a pluralist stance on character reading? Furthermore, is stability even a coherent notion considering the complexities of predicting behaviour—is this because there are no traits to predict behaviour with?

In response to the latter point, in section 3.3 of the introductory chapter, I argued that the situationist attack on the reality of character traits was far from established. Furthermore, the situationist attack does not seem to be an issue for even a vague notion of stability because a notion of stability still seems to track reality in certain ways, where stability allows for effective aggregation:

> even though an individual's level of altruism (or self-control, or honesty) in one kind of situation is not a particularly accurate indicator of his or her level of altruism (or self-control, or honesty) in another kind of situation, people still differ consistently in their overall level as observed across many situations…The effect of aggregating is important, because even though we may have a difficult time in predicting people's behavior in any one specific situation, we can still be rather successful in predicting people's overall patterns of behavior (Ashton, 2013, p. 32).

[50] In addition, despite the language of 'broad' and 'narrow', a middle ground seems conceptually difficult (which is why it is presented as a dilemma): a middle ground between supreme inaccuracy and a failure to predict at all still does not appear to correspond to the reality of trait-based predictions.

Of the former objection, we can escape Westra's dilemma because Westra's critique essentially bottoms out in arguing that pluralists cannot explain trait predictions because representing the situation creates a dilemma unless mindreading is included.[51] Although I defend a pluralist notion of character traits in social cognition, I need not be bound by the difficulties of representing situations as they pertain to their associations with particular traits. This is because my notion of character traits *does* include, contra Andrews, some mental content information.[52] This is detailed in Chapters four and five.

In summary, then, traits being stable is employed because this is what the folk consider traits to be (hence is my target of study); there is also evidence that traits being stable does reflect reality (Ashton 2013). Furthermore, one can avoid the issues of how predictions of behaviour can be made from parsing traits on a classic pluralist account due to the fact that my forthcoming account of traits in social cognition allows for the presence of information about the mental content of others' minds during character reading.

4.2 The Psychological Dimension to Character Traits

Whilst many who work on character traits are concerned with behaviour, as this is what is empirically measurable, there is also a mental dimension that not only grounds the behaviour but also constitutes inward expressions of the traits. This is not an uncommon view—Alvarez noted this of Gilbert Ryle, for example: "Ryle notes, a character trait such as pride is (…) also a disposition to certain 'inner' phenomena such as thinking, judging, reasoning, desiring and feeling in certain ways" (Alvarez 2017, p. 76).[53] My contribution is to give a couple of quick examples to show that this mental/psychological dimension can also be cashed out in dispositions to perceive and experience in certain ways.

4.2.1 Dispositions to Perceive and Experience

In addition to inner phenomena such as those in Ryle's list, there can be perceptual differences between people with different traits. For example, a self-centred person

[51] Although, to be fair to Andrews, Westra does not detail how including mindreading information solves the dilemma. If pressed, though, he would likely note that the consideration of mental states of others is reasoning that goes beyond a trait-behaviour association, where mindreading is resultantly more accurate than character reading.

[52] Though, contra Westra, this information is not present *qua* mindreading.

[53] A psychological dimension to traits also helps explain some people's insistence that certain actions by them are not indicative of their character, such as introverted people pursuing careers as actors.

seems to perceive less about the world around them than very conscientious people; the self-centred person is less likely to perceive the facial expressions or even the language of other people sometimes, whilst the conscientious person will be much more aware.[54]

When it comes to experience, experience includes the mode of presentation for the perception, where for example a happy person might experience a string of unfortunate events as comical, whereas an anxious person might experience them as harrowing. Experience also includes specific mental states. For example, an 'attention-seeking' trait might affect your thoughts about people discussing their achievements at the attention seeker's lifetime-achievement award ceremony. The attention seeker may experience the rudeness of those people much more sharply and may specifically form thoughts about such rudeness. Perhaps they will become agitated because they are not currently the centre of attention. This may result in outward behaviour, but it certainly begins in the mind, and as such constitutes an inner expression of the trait. Additionally, I take it that character traits possess a psychological dimension through affecting our experiences, in that they can help construct our sense of self, though I give no specifics for an account of this. Suffice it to say that understanding yourself as kind—or generous, or analytical, et cetera.—can help you understand the ways you respond to the world around you; perhaps this is why pop-psychology like the Myer's Briggs personality type tests are so popular for this very purpose, despite their lack of empirical validity (Coffield et al. 2004, pp. 48–50).

The point here is that these are but a couple of reasonable examples of what I think is a long list of the ways in which our character traits affect what we perceive and how we perceive, and furthermore influence our thoughts about situations. Any conceptual definition of character traits is going to miss something out if this psychological dimension is not included. This is particularly important when we turn to how trait *attribution* functions in light of our possessed mental state concepts, as I argue in Chapter five. For now, it suffices to note that we can give reasons for character traits as possessing a psychological dimension.

[54] There is a question here as to whether the self-centred person still perceives, but just does not care, or whether they truly do not perceive. Perhaps, however, this is only testable through self-report. That said, Sadeh and Verona (2008) do note that psychopathic people have attentional selection deficits and cognitive control issues relating to their objects of attention, where self-centredness as a trait is considered as a 'Machiavellian egocentricity' sub-domain of psychopathy according to the psychopathic personality inventory (p. 6). As such, the data are at least consistent with my anecdotal claim about the perceptions of the self-centred.

Conclusion

This chapter has given and defended my account of what a character trait is. I identified and answered some metaphysical objections relating to their nature as properties and the inclusion of dispositions, then I outlined my account of traits as tendencies, of their being dispositions plus implicit histories or summaries of manifestation. This distinguishes them from paradigmatic dispositions, which need not manifest in the way that character traits do. I defended this notion of a tendency from various objections. Finally, I gave some reasons to think that character traits have an important psychological dimension, through affecting our perceptions and experience, which will be an important consideration for the relationship of character reading to folk psychology that is considered in later chapters.

Searching for Simulation in Inaccessible Processing

1. Introduction

This thesis is about how we ought to understand character reading in the context of an emerging pluralist folk psychology. I took some time to define the key terms, and to defend the sense of character traits that I am using; now, we need to start considering the socially cognitive mechanisms that underpin character reading. In particular, as established in Chapter one, we have robust notions of theory and simulation as candidates. This chapter sets the scene for an account of the social cognition of character in the following chapter, by due diligence in arguing against a pure simulation account.

As a primer, the small amount of scholarship on this generally claims that simulation is unlikely to have a place in trait attribution—theory-theorist accounts are assumed. For example, if the process of trait attribution is mentioned at all, it is claimed that it is hard to see how trait attribution could be simulative (Andrews 2008, p. 26), or brief arguments against simulation in trait-based reasoning are implied to apply to trait attribution as well (Westra 2018, pp. 1223–1224).[55] The literature is pre-theoretically dismissive of simulation in trait attribution, but we cannot accept such claims until an in-depth analysis has been conducted. Indeed, if character reading is like mindreading in any relevant sense, then a simulationist account of mindreading may apply to trait attribution—I will rule pure simulationism out. Pure simulation cannot result in a trait attribution on any reasonable account adjacent to mindreading; we are left with theory-theorist and hybrid options. In the following chapter, I will then give my own hybrid account and defend it against a theory-theorist point of view.

Section 2 will motivate the examination of potential simulation in reflexive spontaneous trait attributions. It will argue that, given the notion of simulation outlined and in parallel with Alvin Goldman's simulation-focussed hybrid account of mindreading (Goldman 2006), the most plausible places where we might find simulation (without theory) in trait attributions are in those cognitive processes that are—in principle—inaccessible to phenomenal consciousness. Such processes are typically fast, automatic, and immune to conscious interference. In section 3, I will critically analyse how simulation in trait attribution of this kind might function by employing and adapting

[55] This will be covered in more detail in Chapter four.

Goldman's strategies for pure simulation in mindreading. I will show that even if simulation is involved, one cannot impute a trait to someone without the involvement of theoretical inference.

2. Motivating the Search for Simulation

2.1 Initial Plausibility of Simulation in Trait Attribution

This chapter will focus on the notion of *process* simulation, as outlined in Chapter one. This is because character trait attributions do not appear to be *experienced* simulations. Consider how we introspect our own experience. Imagine that I am a child and I know little about the world. I see a person leaking from their eyes and wonder: What are they feeling? Well, I have leaked from my eyes before, and every time that happened it was because I was feeling *this* way with *this* intensity. Note that child-me does not need to know what crying or sadness is. Because I can introspect on my feelings, child-me can make this commonly accurate mental-state attribution, via simulation, because it seems like they are feeling the same way that I feel when I am like that. However, we cannot observe or examine our own character traits like this. Even assuming that I know my own traits, I cannot observe my laziness; I merely observe thoughts and actions that could be said to be indicative of laziness. Indeed, we have blind spots regarding our own traits that other parties can see and reach consensus on (Westra 2019, p. 9). With experienced simulation out of the picture, this leaves process simulation—the search for which is fortunately an empirical project. As reminder, process simulation is the replication of some process that resembles the original process in some way—this might be in the reuse of particular cognitive functions for different purposes, for example.

Here, I outline the sense in which it may be plausible for simulation to be present in making a trait attribution, thereby motivating the search for it. Making a trait attribution involves some cognitive processing. It is coherent that we might simulate some processes of others that are manifestations of traits, particularly manifestations of behaviour or psychology produced by traits. If it is by simulating such processes that a trait is identified (and hence attributed), then a process simulation account of trait attribution is plausible. Understood in this way, the character traits attributed through simulation could be construed as emergent properties which are discernible from particular sets of simulative processes, rather than as discrete entities themselves.[56] To be clear, certain mental states, body language, and actions can all be relevant to making character trait attributions—we should not assume that the sufficient replicative processing of relevant component parts such as these could *never* result in a trait attribution; the sufficient replication of relevant features of the world could construct a painting, or a visual field, for example. Trait simulation is therefore not intuitively implausible, but it still remains to be seen as to whether there is evidence of this.

As such, the search for simulation can be motivated, but I need to address why this chapter specifically looks for simulation in *inaccessible* processing. To do this, I need to introduce and clarify a particular kind of distinction made by Alvin Goldman in his thorough account of simulation in mindreading, from which much of this work is drawn.

2.2 The Low-level/High-level Processing Distinction

Goldman's (2006) book gives a thorough and detailed simulation-weighted hybrid account of mindreading. Of relevance here is his Chapter six, in which he argues that—whilst he endorses a hybrid account for mindreading generally—there is a particular case of *simulation-only* mindreading that can occur. He argues that such simulation occurs in certain 'low-level' cognitive processes, for which the outputs of such processes do not require the input of some theory. His case study is 'face-based emotional recognition': he argues that we can directly mindread the emotions of others from their faces without the intervention of any guiding theoretical knowledge (though there are, of course, additional ways to do it with theory if one so desires). Face-based emotional recognition is an exemplar of what Goldman terms a 'low-level' process. Here, I detail what Goldman means by low levels, in comparison to high levels, but I ultimately cut the issue slightly differently. This is due to some ambiguities and extraneous theoretical commitments regarding the distinction between low and high levels that I note below.

Goldman (2006) describes low-level mindreading processes as "comparatively simple, primitive, automatic, and largely below the level of consciousness" (p. 113). I expand on this feature set by reference to face-based emotional recognition:

[56] I mean 'weak' emergence here—that such 'emerging' properties are derivable from their constitutive components and how their properties interact. If one prefers, one can apply Thomas Nagel's rejection that this does not constitute emergence at all, in which case its use is metaphorical (Nagel 1979, p. 182).

- Simplicity: He refers to ‘emotion types’ as opposed to emotion types plus propositional contents—for example, recognising disgust but not what the disgust is about.
- Primitivity: He refers to a ‘special survival value’ that recognising emotions brings, suggesting that the capacity may be evolutionarily specialised and is implicated as an older phenotypical adaptation than other instances of mindreading.
- Automaticity: Automatic processes, to Goldman, are those that are not potentially, nor intermittently, under any form of intentional guidance or control. This is extrapolated from a quote on p. 133 which notes that “it [a particular form of high-level simulation] is an activity that is potentially and intermittently under intentional guidance or control, whereas low-level mindreading is fully automatic”.
- Being largely below consciousness: Automatic processes are supposedly not potentially under intentional control, but there may be some exceptions. I take it that instances of automatic processes that we are conscious of (but are not under our control) are those that are consciously felt. An example taken from de Vignemont (2009) is that some people feel tactile sensations on their own skin when they watch others being touched (Blakemore et al. 2005).

Firstly, it is not entirely clear that primitivity is required for a generalised notion of low-level phenomena, even if it is apt for face-based emotion recognition. A low-level process might not have a particular special survival value that manifests in an ancient *genetic* adaptation; I will explore natural selection as acting on *culture* in Chapter six. Furthermore, there is some suggestion that Goldman’s terminologies of low- and high-level are specifically hard to distinguish cleanly, particularly in articulating which elements of the differences are necessary, rather than sufficient (de Vignemont 2009). Indeed, such distinctions may be difficult to cut cleanly in general—these distinctions are reminiscent of dual process theories of cognition, which may be plausible but are difficult to clarify where the conceptual boundaries between each system lie (Thompson 2014).[57]

[57] Becoming conceptually clear on the exact boundaries of system 1 and system 2 is a task which goes beyond what I am capable of doing in this thesis; hence, dual process theories of cognition are generally laid aside.

My main concern, though, is that low-level phenomena are purportedly *automatic*—it is their designation as being automatic that causes problems for assessing which processes are low-level and which are not. In discussing Goldman's low-/high-level distinction, de Vignemont (2009) characterises automaticity according to norms of cognitive psychology (Neumann 1984): having no intentional control, being immune to interference, and being not necessarily conscious (p. 461). She assumes that Goldman is using this definition also; Goldman did not make this clear, and so I think this is a fair construal. I suppose that if Goldman intends automaticity to be deviant from the conventions of his discipline, the burden of justification is on him. Therefore, on the assumption that Goldman is using a standard definition of automaticity (given the lack of evidence to the contrary), instances of otherwise-automatic phenomena that *can* be interfered with are potential points of challenge for this low-level/high-level distinction.

Of course, what counts as being 'interfered with' is not clean-cut either. The example usually given is the Stroop effect (Stroop 1935). This is where you are tasked to assess the colour of a word, but you are unable to stop yourself from reading the word and thus invoking processing time difficulties if the word itself is a different colour than the colour of the word (for example, the word is 'red', but its lettering is coloured green). This seems fine, but I think the crucial point is whether contextual knowledge that affects processing will count as 'being interfered with'. There are processes that function without our conscious awareness, and that are not capable of being intentionally controlled, that can be nonetheless interfered with by our contextual beliefs —are they no longer automatic? Socially cognitive phenomena, for example, such as the purportedly 'low-level' face-based emotional recognition, are cross-culturally variable based on our cultural epistemologies (Jack et al. 2012). In her paper, de Vignemont mentions that 'mirror empathy' is not immune to interference despite meeting other desiderata for automaticity (p. 5), and the example of otherwise automatic trait attribution that I will discuss below can also be interfered with by contextual knowledge, such as whether one's interlocutor is a stranger or a friend (Rim, Uleman and Trope 2009). Finally, there is a big debate on whether contextual knowledge interferes with our perception in the cognitive penetrability debates (Stokes 2013), despite perception's basic processes being unable to be consciously controlled. As such, I take it that low-level phenomena being automatic, as Goldman suggests, introduces too much uncertainty for the categorisation of the relevant subjects of discussion in the exacting manner that is desired.

My task is therefore to capture the spirit of what 'low-level' entails without incurring objections based on these extra theoretical commitments of survival value and automaticity. Suppose we understand the processing as it cuts across the following familiar divides: Whether such processing is, in principle, accessible or inaccessible to consciousness. A process that is in-principle accessible to (phenomenal) consciousness is one which might not be a feature of awareness in one's experience, but one that may come into one's awareness under certain circumstances. Such circumstances might be introspection or being made aware of the process to attempt to inhibit it, for example.

Theory and simulation can be in-principle phenomenally conscious in such cases as explicit reasoning; for example one may directly theorise or aim to simulate for the purposes of reasoning ('I wonder how I would feel if that happened to me? I will try and see what follows'). This is the subject of Chapter four, but what Goldman calls a low-level process, we will call an in-principle inaccessible-to-consciousness process. Such phenomena are unable to be called into phenomenal consciousness, whether by choice or force. Certainly, Goldman would agree that this moniker fits for face-based emotional recognition, seeing as he thinks that this occurs as a direct result of mirror neuron matching—a phenomenon that is certainly in-principle inaccessible to consciousness. Indeed, this appears to capture the spirit of the automaticity talk above, in which the given distinction is that a low-level process is fully automatic, whereas a high-level process admits some potential intentional control. In opting for inaccessibility to consciousness, as opposed to talk of what constitutes low-level, this also means that—when comparing to trait attribution—we need not be distracted by tangents invoked by Goldman's notion about, for example, primitivity and its supposed special survival value, or the fact that whether automatic processes are immune to interference is contested.

Having clarified Goldman's use of a low-level/high-level distinction between kinds of cognitive processing and having outlined how my own work relates to this distinction, we now need to motivate simulation in trait attribution by way of a target that can be analysed. Is there such a thing as a trait attribution that functions in a way that is inaccessible to consciousness? Yes. For this, I reintroduce spontaneous trait inferences as the target phenomenon, given that its processing and output is unable to be brought to phenomenal consciousness, before moving to the case study of attempting to use Goldman's simulative mindreading account on spontaneous trait inferences.

2.3 Spontaneous Trait Inferences as Inaccessible Processing

Spontaneous trait inferences were introduced briefly in the introductory chapter. As a reminder, when we meet someone, we quickly, unconsciously, and without our intentional control attribute traits to them.[58]

These are known to be unconscious because in Winter and Uleman (1984)'s paper on spontaneous trait inferences, after completing both experiments testing for the phenomenon, they asked participants to estimate the percentage of the time that participants considered personality in the judgments they made during the trials. As a reminder, the experiments themselves were memory recall tasks that evidenced the spontaneous trait inferences through quicker recall of words that implied traits that were previously attributed (3.2.1 in the introductory Chapter). As Winter and Uleman note:

> Although demand characteristics would predict that subjects in such situations would strive to be agreeable and confirm the experimenter's suggestions, most subjects regretfully reported having made no such judgments at all. Even after the debriefing, some did not believe they had made trait inferences and were greatly surprised by evidence supplied by their own recall sheets that trait cues had actually been effective in promoting their recall (p. 245).

It should also be noted that such attributions lack intentional control because that is the nature of the 'spontaneous' designation, that is, such attributions occur unconsciously and without intentional control. In cases where intentional control can be exerted, these are not spontaneous trait inferences, these are intentional trait inferences.

This, of course, raises the question as to whether we ought to consider that spontaneous trait inferences and intentional trait inferences are really two different kinds of trait attribution at all. Indeed, Willis and Todorov (2006) found that intentional trait inferences could occur in response to presentation of human faces as fast as 100ms after exposure, so even intentional trait inferences can be as fast as the supposedly automatic spontaneous trait inferences, which might cast doubt on the difference

[58] Poignantly for their nature as lacking intentional control, Levordashka and Utz (2017) found that not only did evidence for spontaneous trait inferences occur when viewing social media status updates, but when scrolling through many updates, it appeared that sometimes trait attributions were mis-associated with other profiles (pp. 97–98).

between spontaneous (low-level adjacent) and intentional (high-level adjacent) processes of trait attribution.[59]

Firstly, it should be noted that I have not made any commitments of speed between low- and high- level processing—a requirement for such a commitment is usually levelled against dual process theories (Evans 2012, p. 22). As noted above, I am not subscribing to a dual process account for character reading, though even if I *ought* to have done, I could accept both of the following outcomes. Either 'fast' processing phenomena need to be governed by system 1, or they can be governed by a system 2 that operates as fast as system 1 in that instance. If they need to be system 1 in order to be fast, then this is coherent with a claim that practice, expertise, and the introduction of heuristics for certain processes can enable system 1 to govern that process where it could not previously. Hence, fast trait attributions could be made intentionally with system 1. If, on the other hand, system 2 can be just as fast, the same 'expertise' claim is still coherent: through practice and applying heuristics, system 2 can perform faster for given tasks—this is argued for in Evans (2010, chap. 4), for example.[60] Given the acceptability of both outcomes, processing speed cannot distinguish between types of trait attribution.[61]

Regardless, there are other reasons for thinking that spontaneous trait inferences and intentional trait inferences are different kinds of trait attribution. Firstly, Van Duynslaeger, Van Overwalle, and Verstraeten (2007) discovered that intentional trait inferences activate the brain area of the medial prefrontal cortex strongly, whereas spontaneous trait inferences activate the temporoparietal junctions strongly (pp. 181-182). This suggests different kinds of trait attribution by the differing neural locale of their processing. This is not decisive, of course, so I will say more: we can appeal to a difference in kind of trait attribution by certain differences between them: in (a) the

[59] Note that these spontaneous trait inferences are not automatic in the strict sense discussed in the previous subsection: Uleman, Newman, and Winter (1992) found that cognitive capacity could interfere with whether spontaneous trait inferences were made. Hence, spontaneous trait inferences are better conceptualised under my distinction of whether such inferences are accessible to consciousness (for the purposes of this discussion).

[60] That said, there is scepticism as to why we would need a system 1 if system 2 can be just as fast. This is expressed in Thompson (2014), noted above, and is at least one motivating factor for my not committing to dual process theories in this dissertation.

[61] Therefore it is not significant that Willis and Todorov found, for example, that "judgments made after 100-ms exposure to a face were highly correlated with judgments made in the absence of time constraints" (p. 596).

'catalyst' for the inference, (b) the extent of the monitoring, and (c) the use of the inference (Ferreira et al. 2012, p. 2).

a) 'Catalyst' is Ferreia et al.'s terminology and I am not sure the word specifically applies. The relevant point is that trait inferences are made in response to certain phenomena, and that this is different for spontaneous over intentional inferences. Spontaneous trait inferences are made in response to perception, they note, whilst intentional trait inferences are the results of intentions to form impressions.

b) There is also a difference in monitoring. Unlike intentional trait inferences, for spontaneous trait inferences, there is no conscious monitoring of the outcomes of the trait attributions made (given as they are only notable when discovered in memory recall tasks). This is not to say that spontaneous trait inferences are not updated in response to new information, but that such updating is not consciously monitored. As such, there is a meta-monitoring phenomenon that constitutes part of the difference between the two kinds of attribution.

c) Finally, there is a difference in goals between the two types. Intentional inferences are goal-directed, whether it be towards impression formation, or prediction and explanation of events, for example. Spontaneous trait inferences are, on the other hand, not goal-directed. This is not to say that they are not sensitive to goals—whether spontaneous trait inferences are made at all can be interfered with according to a person's goals, where for example memorising sentences with the goal of ignoring their meaning in fact reduced, but did not eliminate spontaneous trait inferences (Uleman and Moskowitz, 1994, p. 494). The point here is that the inference itself is not directed towards a specific goal of the agent, unlike intentional inferences.

As such, I take it that with no issues arising from processing speed, evidence of a differing neural locale, and differences according to functioning and purpose, we can usefully describe spontaneous and intentional trait inferences as two different kinds of trait attribution.

Relevantly for the overall discussion on Goldman, spontaneous trait inferences are in-principle inaccessible to consciousness, compared to the kind of trait attribution I make when I deliberate on someone's behaviour with the explicit goal of attributing a trait. This is noted as a result of, for example, those participants in Winter and Uleman's (1984) study who struggled to believe that they had made spontaneous trait inferences, even after having seen the memory recall data—if such attributions were accessible to consciousness, it is surprising that those participants primed with the goal to bring such attributions to consciousness could not do so. As such, spontaneous trait inferences are the relevant trait attributive phenomena for assessing simulation in trait attribution within the Goldman-centred dialectic of this chapter.

Returning to spontaneous trait inferences with regards to assessing the data on whether they are processed according to theoretical or simulative means, unfortunately, investigations of such processes are currently scant. There is *some* empirical work on the neural basis of spontaneous trait inferences, in which for example, as noted above, Van Duynslaeger, Van Overwalle and Verstraeten (2007) showed that intentional trait inferences activate the medial prefrontal cortex strongly, whilst spontaneous trait inferences activate the temporoparietal junctions strongly (pp. 181-182). This has the potential to say something about *intentional* trait inferences, since the medial prefrontal cortex is also involved in self-reflection; hence a discussion of simulation might be introduced (e.g. Mitchell, Macrae and Banaji, 2004, p. 4915). However, this is still fairly speculative and lacking in detail, whilst also saying little to nothing about the target phenomena of spontaneous trait inferences.

Van Duynslaeger et al. (2008) also analysed electroencephalographical components (measuring brainwaves) of spontaneous trait inferences, but their conclusion is merely that because people are sensitive to conflicting trait information, it must have been the case that they had spontaneously attributed the trait initially; this consclusion does not appear to be able to adjudicate one way or the other for theory or simulation.

To reiterate, I mention the neural data above only to note that, whilst there is empirical work on spontaneous trait inferences, the relevance to theory or simulation (from what I could find) appears to be lacking.[62] Beyond this, drawing conclusions about theory-theory and simulation theory from neural data has been historically difficult, at least in the case of mindreading. This is because, as (Apperly 2008) argued, despite the optimism noted in Stich and Nichols (1997), studies up until Apperly's publication in 2008 had failed to adjudicate between the two views. This was due to methodological issues surrounding whether the selected studies discussed truly relevant mindreading content (beliefs and desires), and whether they used an appropriate notion of the self

[62] Though, a full review needs to be conducted to assess that claim.

for simulation (pp. 274-276). Furthermore, Apperly raised the concern that a legitimate neural test of theory-theory against simulation theory has the following issues:

> [it] depends upon a scientific account of where, how and when mental states such as beliefs and desires are formulated in the 1st person case…Indeed if future work fails to solve the more difficult problem of identifying conditions under which we can be sure that a participant is in a current state of believing, intending or desiring then testing ST [simulation theory] against TT [theory-theory] in this way will be impossible (p. 276).

If such testing is still contentious for mindreading, I see no reason why the same would not be said of trait attribution, given that comparatively less research has been done in total in the trait domain. That said, Apperly is optimistic that we can get a lot of interesting data from such studies that have already been done on mindreading, and some issues with methodology could be solved, in principle (pp. 279-280).

Regardless of the neural data being currently unhelpful for testing theory-theory against simulation theory, in summary: spontaneous trait inferences are the target phenomenon here because they are in-principle inaccessible-to-consciousness trait attributions. Hence, if Goldman's thorough account of simulation in 'low-level' mindreading is in any way applicable to some account of trait attribution, spontaneous trait inferences are the relevant phenomena. As such, I move back to a discussion of Goldman's account of pure process simulation in mindreading. The next section details how such processes are supposed to work, and I evaluate them for applicability to spontaneous trait inferences.

3. Simulation in Inaccessible Processing

3.1 Goldman on the Evidence for Pure Simulation

Here, I outline the four best pure simulation inaccessible-processing hypotheses that Goldman (2006) considered for face-based emotional recognition mindreading.[63] After a brief clarification on Goldman's preferred method and why he rejects the other three, I will assess these simulative mindreading processes in relation to trait attribution and argue that none of these options allow for pure simulative trait attribution.

[63] Each hypothesis was generated from extensive empirical research; I take these options to be sufficiently rigorous, if not exhaustive.

1. Generate and Test

   You see a target's 'facial display', you hypothesise an emotion as the cause, then you generate a 'facsimile' in your own 'system'. For example, you see a frown, you hypothesise 'angry', then you instantiate angry in yourself; the resulting frown is confirmed by the simulation. The hypothesis is confirmed and "imputed to the target" (p. 125).[64]

2. Reverse Simulation

   You 'covertly' imitate the facial expression you see through miniscule musculature movements, where these imitations produce "traces of the relevant emotion" (p. 127): whatever emotion is produced in trace quantities is attributed to the person that triggered the simulation.

3. Reverse Simulation with an 'As-If' Loop

   After the presentation of a visual representation of a facial expression, an 'as-if loop' in the somatosensory cortex is created that bypasses facial musculature in order to create a sensation of what it would feel like to have that facial expression. After the activation of an emotion which is relevant to the felt sensation, there is a recognition of the emotion and an imputation to the target. This variant of reverse simulation exists because brain lesion research demonstrated that the recognition of facial emotion requires the 'integrity' of the right somatosensory cortices (Adolphs et al. 2000).

4. Unmediated Resonance (Mirroring)

   "Perception of target's face 'directly' triggers (subthreshold) activation of the same neural substrate of the emotion in question" (p. 127). This draws on evidence of mirroring for the disgust emotion (Wicker et al. 2003, p. 661) and the mirror neuron systems literature generally (Rizzolatti et al. 1996).

Goldman favours the fourth option and invests much time into the discussion of mirror systems. Mirror neurons (of the mirror systems) are neurons that fire both when one acts and also when one observes another acting in a similar manner. For example, raising your hand and watching someone else raise a hand will both fire mirror neurons. Mirror neurons were first discovered in monkeys (Gallese et al. 1996; Rizzolatti et al.

[64] Talk of hypothesising does not imply conscious reasoning–it refers to the process by which this automatic matching occurs, and thus still remains a pure simulation. The relevant question, of course, regards how this emotion is selected; see 3.4.1.

1996), then purportedly—though this is still contested— in humans (Rizzolatti, Fogassi, and Gallese 2001). Goldman notes:

> These neurons seem to constitute an execution/observation 'matching' system, or 'resonance' system. Certain neural activity in an observer resonates with the neural activity in an observed actor. Each family of mirror neurons comprises the substrate of a distinctive type of (nonconscious) mental representation, something like a plan to achieve a certain behavioral goal (grasping or tearing, for example). In the case of the observer, however, the plan is not executed (Goldman 2006, p. 134).

It is on this basis that Goldman spends much time showing how mirror systems and mindreading relate. For now, I put aside the issue in favour of briefly explaining why Goldman rejects the other three possibilities, before moving to my analysis of them in relation to character trait attribution; their failure in relation to trait attribution sometimes requires a different approach. We will see later that the mirror neuron hypothesis cannot work for trait attribution, as it encounters the same problem that other methods do when trying to use inaccessible process simulations to attribute traits.

Goldman's concern for 'generate and test' is in how to choose the generated emotion to test, as a random selection of the six basic emotions (anger, disgust, fear, happiness, surprise, sadness) would be too slow. Theoretical guidance could do it, but then it is not purely simulative (p. 129). He has multiple listed concerns for 'reverse simulation', based on rival explanations of the data, but the data that does not support the truth of this hypothesis is crucial. Hess and Blairy (2001) showed that the "successful mimicry [of a perceived facial expression] did not correlate with accuracy in facial recognition, suggesting that facial mimicry may accompany but not actually facilitate recognition" (Goldman 2006, p. 130). Regarding 'reverse simulation with an as-if loop', Goldman notes that the impairment of specific emotion recognition did not present as differences in lesion overlap between subjects, so "it is not clear that activation in this region [of the brain] is specific enough to recognise one emotion as contrasted with others" (p. 131).

We have some context for these possible methods of inaccessible simulation for mindreading, then, but now I move to my own evaluations as they might pertain to trait attribution. I will tackle these options in a different order than Goldman has, i.e., in the order of the plausibility of their application to trait attribution. After all, face-based emotional recognition is a very specific type of mindreading, and the evidence for these methods is—in some cases—specific to faces and emotions; we would not expect all of the above methods to be easily and obviously applicable to trait attribution. I take the reverse-simulation methods first, then mirroring. In the discussion of mirror neurons, I highlight a certain 'combination' problem that plagues it as well as the final, most plausible, option: 'generate and test'.

3.2 Applicability to Trait Attribution: Reverse Simulation and Reverse Simulation with an 'As-If' Loop

Goldman only considers this option because we know that we often feel happier after smiling, but generally backwards mindreading is not possible. Applied to traits, we might think that acting like a good person may indeed make someone feel like they are a good person, but it is nonsensical that imitating someone else's being a good person then follows to you generating a belief that they are a good person. This is because one must already have this belief in order to successfully imitate them being a good person.[65] In addition, imitating someone being a certain way seems to imply much more than the inaccessible processing occurring in facial mimicry. Indeed, I discuss taking on someone else's likeness in the next chapter. Regardless, Goldman notes that "the standard forward directionality of mental processes precludes the possibility that these processes can be utilized in the opposite direction" (p. 125), where face-based emotional recognition would have been an exception. As such, I share Goldman's scepticism of this variant.

Secondly, the as-if loop version incorporates empirical data specifically in relation to the neural correlates of face-based emotional recognition; it is difficult to make the case for such a point applying to trait attributions. This is because there does not appear to be evidence that trait attributions share a *reliance* on the right somatosensory cortex for their functioning. Indeed, trait attribution (inclusive of spontaneous trait inferences) appears to use completely different parts of the brain, though which parts specifically there is not a complete consensus on. For example, some studies bundle trait attribution and self-reflection together to suggest that the medial prefrontal cortex is mostly involved (Johnson et al., 2002; Kelley et al., 2002; Schmitz, Kawahara-Baccus and Johnson, 2004), and a meta-analysis also suggests the medial prefrontal cortex as being

[65] One might wonder if one could imitate their behaviour under a different description and come to a trait attribution this way. This might be possible, but then it would certainly not be an imitation of them being a good person under *that* description; imitation of behaviour understood under *that* description is required for *reverse* simulation.

crucially important for trait attributions (Van Overwalle, 2009, pp. 846–848).[66] Other studies implicate the right temporoparietal junction for intentional trait inferences (Van Duynslaeger, Van Overwalle, and Verstraeten 2007), and some studies suggest that even the posterior cerebellum may be involved in some aspects of trait attributive processing, in "supporting an active process of sequencing trait-implying actions" (Pu et al. 2020, abstract).

The closest one might come to a neural overlap between emotion and trait attribution appears to be in the function of a differing neural location than the right somatosensory cortex. In one study, an area of the *dorso*medial prefrontal cortex was activated when character impressions were made of others (Mitchell, Macrae and Banaji 2004), where the dorsomedial prefrontal cortex is implicated in generating and regulating emotion (Kober et al. 2008, p. 1022).

However, a different study concluded that there were (at the least) partially dissociable neural systems between emotion attribution and trait attribution: Heberlein et al. (2004) investigated the relationship between emotional attribution and trait attribution to point-light walkers, and found that the attributive performance was dissociable based on legions in the right somatosensory cortex (for emotions) and in the left temporoparietal junction (for trait attributions). As such, whilst the neural locale of trait attribution is somewhat contested (though the medial prefrontal cortex seems like a good bet), none of the studies that I could find implicated the right somatosensory cortex, let alone a specific reliance on it; Heberlein et al.'s study even suggests against such a reliance due to the dissociations evidenced through lesions.

Because of this apparent difference in processing locale, a simulative neural reuse argument appears to be off the table for an as-if loop argument. That said, this claim may depend on some assumptions about simulation and mirroring, namely that simulative neural reuse for mirroring requires the activation of those similarly physically located neural processes of the target function to be simulated. Our knowledge of mirror mechanisms is still incomplete, particularly in integrating different levels of description of action and emotion processing (Rizzolatti and Sinigaglia 2016, p. 8), so further review is likely needed. For now, though, I eschew further discussion of these

[66] Notably, Mitchell, Banaji and Macrae (2005) understand the complexities surrounding disambiguation of these factors, such as whether trait attribution plus self-reflection lets us infer anything about trait attributions alone.

variants because they appear implausible even in the face-based emotional recognition case, let alone being adaptable to trait attribution.

3.3 Applicability to Trait Attribution: Unmediated Resonance

As I previously noted, Goldman builds his case for pure simulation in mindreading on mirror systems in which, as they pertain to face-based emotional recognition, "perception of target's face 'directly' triggers (subthreshold) activation of the same neural substrate of the emotion in question" (p. 127). The first issue in the applicability to trait attribution is that it seems that the direct triggering of the same neural substrate of an emotion in question, as observed, will not work for traits, as traits are not states.[67] Character traits are not represented in the same way that mental states are, hence the previous issues with introspecting them (in comparison to mental states).

Even granting that there are neural substrates for emotions, it is hard to see how there might be such physical substrates for particular traits, given that traits themselves are dispositions plus histories. One might suppose that a detailed look at the mechanics of mirror neurons might yield a more satisfactory account, but I do not think we need to take the time or space to do so. This is because the 'unmediated resonance' account encounters another particular problem—independent of neuron activity—that I introduce in the following subsection. Note that this issue, the 'combination problem', applies to both the unmediated resonance account and the following 'generate and test' account.

3.3.1 The Combination Problem

If resonance is unmediated, in the sense that direct perception activates some emotion, we run into a combination problem. The problem borrows its name from the structurally similar problem in debates over panpsychism (Chalmers 2016). In that debate, the problem regards how there does not seem to be any room for an account to combine perspectives of smaller conscious units into our unified perspectives—panpsychism supposedly does not have the ability to explain the apparently nonsensical destruction of multiple perspectives when combined into a larger one. The combination problem for unmediated resonance is that the account does not have the resources to explain how resonance of the markers of personality then results in a trait attribution without further mediation. The point is that there is no room for theoretical inference

[67] Dispositions might be states, but traits are not mere dispositions.

in such an account because this would constitute a mediation after the resonance, and the simulation is supposed to be the resonance itself. As such, how the resonance of the markers of personality then somehow combines into a trait attribution without a further step constitutes the combination problem.

There is a second related issue here: Emma Borg has noted of mirror neuron activity that bodily movement, or "brute kinematics" as she terms it, clearly underdetermine action (Borg 2007, 2013). One might see someone grasp a cup (and thereby activate mirror neurons) but the goal might be to drink from it, or to throw it. The point applies to traits as well—in mirroring the markers of personality one underdetermines the particular traits that might be attributed. Many behaviours of grumpy and morose people overlap, for example. As such, even if one can give an answer as to *how* we might move from unmediated resonance to trait attribution without further mediation, then one encounters a further problem of how to combine bodily movement into trait attributions when bodily movements underdetermine traits. Of course, theoretical inference could plug this gap, but once again this would reject the purportedly purely simulative nature of mirror resonance.

This foreshadows the conclusion to this chapter, that socially cognitive accounts of character reading cannot be simulationist alone. Firstly, however, there is one more of Goldman's methods yet to assess, and secondly it is not yet clear why a theory-theorist account is not preferable, or why simulation ought to play a role despite the difficulties I am presenting here. I turn to these issues now.

3.4 Applicability to Trait Attribution: Generate and Test

Suppose we see a facial expression. Applied to mindreading, this method would see us generate a hypothesis of the emotion that the facial expression reveals, let a facsimile of the emotion run in ourselves, check it for a match, and then (upon matching) attribute the emotion to the person. The issue we have been running into thus far is that a lot of these methods do not directly apply to traits. I think that this is true here too, so I want to be generous and suppose that the generate-and-test methodology might apply to more than merely face-based emotional recognition as it pertains to traits. This is on the supposition that features that are relevant inputs for the consideration of trait attributions include more than just faces. Suppose that, instead of merely a face with its perceived emotions, we allow informational input for the testing methodology to come from the whole person in their environment. For example, take a man with a pencil moustache, in a grey suit with a grey tie, sitting at the head of a long table. His expression is a mild frown but presents otherwise neutral features. For argument's sake, suppose that we ultimately attribute the trait of 'austere' to this individual. Given that we have specified our entire visual field as input, rather than merely that part which includes the face, we can now assess how the 'generate and test' strategy might function for trait attribution.

I should note that I spend a little more time on this option than the previous reverse simulation and unmediated resonance accounts. This is because I think that its discussion will highlight that simulation is likely to be involved in low-level trait attribution, just not by itself; hence, the resulting hybrid account of character reading explains the phenomenon better than a mere theory account would do.

3.4.1 How to Select for Testing

The first issue we encounter for the generation and testing of our trait hypothesis is in how we choose the trait to test. Goldman was unhappy with 'mere chance' for the selection of one of the six basic emotions due to the perceived unacceptable processing time this would invoke. This is because the wrong emotions might be tested up to five times before the correct one is attributed. I think it is clear that this problem also applies to traits. For individual traits, of which there seem to be many more than emotions, the selection problem appears to be bigger. Even if we only restrict testing to the basic dimensions of individual personality differences, as per the five-factor model discussed in the introductory chapter, then the wrong underlying dimension might be tested up to four times—this still appears to invoke a similarly unacceptable processing time, given the 100ms attribution time noted for spontaneous trait inferences.

Goldman suggests, though, that a *theory* could guide the choice of emotions for testing. It is not detailed, but I suppose a theory could guide the choice by simple rules such as, for example, not selecting a negative emotion if the facial display has upturned corners of the mouth (and vice versa). This would at least reduce the random selection by an appreciable degree. This would also make the strategy a hybrid one instead of pure simulation, but Goldman is not worried by this, as he sees generate and test as being poorly suited for *emotion* simulation. If, on the generate-and-test strategy, a theory of traits *did* guide the generation of a trait for testing, pure simulation is obviously out. However, there is a concern that simulation could not be involved *at all*. We might ask: Why bother confirming the match with simulation?

Take the emotion case as an example. You see the frown, you hypothesise 'upset', you generate a facsimile of being upset in yourself (where this involves the facsimile emotion running 'its typical course', so here at least generating a frown), then you match the facsimile process to being 'upset' based on its similarities, and once that is confirmed you can attribute 'upset' to the target.[68] For traits though, information beyond what is proprioceptively imitable may well be the *most salient* to the generation of the trait hypothesis. By 'proprioceptively imitable', I mean those features of the target that can be simulated by your own body's movements and sensations, such as the almost-indiscernible minute muscular changes one might undergo in running the facsimile of a frown. The man's bland suit, bland tie, unfun moustache, the fact that he is sat at the head of a long table…all of these may be the things that contribute to matching the encounter to the trait 'austerity', perhaps even more so than his particularly neutral facial expression.[69] These are not what one can confirm a match for with a proprioceptive simulation; if theory guides the generation of the trait hypothesis, it is not clear that you would be able to let a trait simulation run its typical course, as you might with emotions. In which case, you must either reject 'generate and test' or accept that the guidance of hypothesis testing is based on a full theory of traits, which of course then negates the need for any 'hypothesis testing' simulation.

Thus, there is a problem for this account in selection for testing, but I want to grant some imagined solution to this (for now) because there may still be some simulations that are relevant to trait attributions—in particular, those simulations of behaviour that *are* proprioceptive imitations. If such simulations are in-principle inaccessible to consciousness, then they might count as relevant phenomena for spontaneous trait inferences. I will highlight a couple of cases of these, but I will argue that even in these cases, such simulations are not in-principle inaccessible to consciousness.

3.4.2 Simulation in Physical Imitation

I return to the austere man. It seems reasonable to suppose that our proprioceptive systems may respond not only to *facial* expressions, by creating facsimiles, but also to other *physically imitable* attributes, such as posture (Shockley, Santana, and Fowler 2003)

[68] Facsimiles of being upset running their typical course might (in theory) refer to actions like lashing out or crying, but Goldman is only interested in processes like the minute facial musculature changes invoked in the facsimile of the frown.

[69] One might wonder why clothing contributes anything towards the property of austerity possessed by its wearer. On my account, their choice to wear certain clothing (and even their choice of home décor) may count as histories of disposition manifestation.

or speech pattern (Delvaux and Soquet 2007; Kappes et al. 2009; Babel 2012). Certainly, we imitate all manners of behaviours of others, known commonly as the chameleon effect (Chartrand and Bargh 1999). Indeed, Lakin et al. (2003) suggest that the fact that there is a correlation between the mimicry of others and successful rapport building serves an evolutionary function by aiding in building interpersonal relationships. Judging character through these simulations that contribute to spontaneous trait inferences would at least fit such a story, where even such a 'primitive' backdrop with 'special survival value' for these simulations would also suit Goldman's notion of low-level phenomena. Indeed, there is evidence of these simulations as being highly useful. For example, Pickering and Garrod (2007), and Knoblich and Sebanz (2016), make a case for the simulation of perceived actions aiding in the prediction of upcoming speech sounds. Perhaps, then, if the contents of the set of trait-relevant behaviours for a given trait can be simulated through these processes, a case for purely simulative trait attribution could retain life.

However, such simulations fail to meet the bar for low-level processes, let alone inaccessibility to consciousness. Garnier, Lamalle, and Sato (2013) found that whilst subjects unknowingly imitated the pitch of vowel sounds that they heard (when they were told to repeat the vowel that they heard); the effect was inhibitable once the subjects were made aware of this, and one participant even overcompensated for the effect. The point here is that if an effect is inhibitable, then the subject is able to exert an element of intentional control over the process; hence, at least some part of the process is phenomenally conscious, however weakly it may be. As such, inhibitable effects fail the lack of intentional control criterion for low-level processes and the inaccessibility-to-consciousness criterion that I set for such phenomena. Whilst Garnier, Lamalle and Sato did note that "phonetic convergence [mimicking others' speech patterns] (…) may primarily be the consequence of an automatic process of sensorimotor recalibration" (p. 12), Leighton et al. (2010) found that social attitudes themselves modulated these kinds of social imitation. The point being that it becomes hard to characterise trait attributions that result from such simulative processes as 'low-level' simulations if they can be consciously inhibited and affected by one's pre-existing social attitudes towards them. Likewise, whilst the processes responsible for mimicking exhibited in the chameleon effect may not be accessible to consciousness, the output is consciously inhibitable.[70]

One might object that these imitations *are* automatic in some relevant sense, but we have merely learned as adults to inhibit them. For example, Brass et al. (2003) and Spengler, von Cramon and Brass (2010) presented studies demonstrating that patients with certain brain lesions appear to lose control of inhibitory mechanisms for imitation: they constantly repeat and imitate experimental controllers, suggesting that the inhibition of automatic processes is a cognitive 'add-on'. This claim is far from confirmed, as—for example—Garnier, Lamalle and Sato (2013) note of their studies that "at the neural level, no additional region or network, out of the typical networks of speech production and perception, appeared to be specifically involved in imitation inhibition" (p. 12). However, even supposing the truth of these conclusions, the point remains that the outputs of such imitation processes (whilst primitive) are *no longer* automatic if they are normally inhibited. Neither are they automatic in the sense that they run independently of the influence (interference) of more complex states, such as those with propositional contents (like social attitudes).

Whilst such physical imitations of others are in-principle accessible to consciousness, this might present a further avenue of enquiry for spontaneous trait inferences if indeed simulation plays a part in trait attributions through such imitations. The process by which spontaneous trait inferences are made is currently obscure, as I noted with regard to the lack of data on the issue. On the reasonable supposition that certain physical features of others are relevant to the attribution of a particular trait, then their proprioceptive simulation may well be part of the picture for making the attribution, as highlighted by the consideration of potential face-based emotional recognition. It may be, though, that some of these imitations are accessible to consciousness whilst others are not—whilst that makes all the difference here in demonstrating that the generate-and-test method is likely not an appropriate account of in-principle inaccessible-to-consciousness trait attributions, this does not imply that in-principle *accessible*-to-consciousness trait attributions cannot make use of such simulations. That said, there are currently no data on this and thus it is not a claim I will endorse here.

[70] Ironically, one can imitate the chameleon effect, where doing so is often used as a deliberate technique by mentalists, confidence gurus, and scam artists in order to garner trust.

3.4.3 The Combination Problem Revisited

With that aside concluded, we can return to the issue of the generate-and-test methodology. Goldman rejected the generate-and-test method on the basis that (for emotions at least) theoretical inference seemed to be required in order to make a reasonably accurate emotional attribution that was not arbitrarily assigned. This might likewise be the case for trait attribution, but we cannot *prima facie* reject that the function of the imitations discussed above are to aid in such accuracy. On reflection, though, the generate-and-test method does seem to have one final nail in its coffin, because any accuracy of the trait attributions selected by the imitations would still need to be explained by a theory; there appears to be no way to combine such imitations into a trait attribution without it. As such, we return to the combination problem.

Individually, we might produce a plethora of low-level simulations, but none are individually sufficient for a trait attribution. The information needs to be considered holistically, so how might one account for this combination? There are three options that I foresee. Firstly, there is mere theoretical guidance. Secondly, the generate-and-test strategy would utilise higher-level processes. Thirdly, the presence of enough of these trait markers triggers some kind of direct match to a trait. Regarding the first option, we are of course denying pure simulation. With the second option, the purpose of positing low-level processing is defeated if it needs to rely on high-level processing. Of the third option, we might ask how such a match could be achieved without theory, but a response of unmediated resonance unfortunately takes us back to the combination problem.

In short, it does not appear as if a 'generate-and-test' strategy is a feasible account of purely simulative trait simulation, even taking so much for granted. The generation seems to require a theory of traits, thereby nullifying the need for a 'test' function. Simulations that could feasibly be involved in trait attributions, such as face-based emotional recognition, suffer from being insufficient alone for the attribution of a trait. Once we consider multiple proprioceptive simulations helping to pinpoint a trait, many plausible inclusions fail to be simplistic or automatic as the special case of low-level mindreading would require. Finally, assessing the collection of simulations also seems to require a theory of traits. What *is* usefully shown, though, is that the generate-and-test strategy provides a basis for the way in which simulations (of proprioceptively imitable representations) might feature in trait attribution (alongside theory)—nothing discussed thus far denies that simulation occurs in trait attribution; it merely denies that it can occur *purely* simulatively. Given the analysis of this strategy, we can see how useful simulations such as face-based emotional recognition may be in quickly providing trait-relevant information to be checked against a theory of traits.

3.5 Relating Trait Attribution and Mindreading Through Simulation

In this chapter, the methodology for the assessment of whether there might be pure simulation in trait attribution was accomplished by a comparison to a detailed account of the same in mindreading. Clearly, there is some kind of connection between mindreading and character reading, given the attempts to either connect them via a singular account (Westra 2018) or to distance them based on particular attributive criteria (Andrews 2008). What we learned about the relationship between the two in this chapter is that, at least in one case, a mindreading account is not transposable to a trait-attributive account. This suggests a more distinctive difference between the two than some might assume, but it also complicates the relationship between them; my full account of the relationship between mindreading and character reading is given mainly in Chapter five, but also Chapter six.

Conclusion

What have we learned? Goldman's work is the most detailed account of pure simulation in mindreading, given its rich theoretical structure that is informed by a plethora of empirical data. However, it seems that none of its posited strategies will be applicable to trait attribution. We can be reasonably confident that no pure simulation account of trait attribution is possible, but this leaves open the question of whether a theory or hybrid account is the better option. The takeaway of this chapter is not to imply a complete rejection of simulation, given that simulation in the generate-and-test methodology was only problematic because it could not be achieved without theory, not because it could not be used at all. As such, I now move to present what I think is the precise role of simulation in a hybrid account of character reading.

Character Reading

1. Introduction

This chapter gives my account of character reading, in particular the relationship between simulation and reasoning with character traits. Building on the lessons learned in Chapter three, I will show that despite existing work claiming that character trait attribution and reasoning are wholly based on tacit knowledge of a theory of the concepts and psychology involved, processes of simulation can be involved. This is in addition to theoretical reasoning, such that future work on character traits in social cognition ought to be couched in a hybrid theory that highlights the flexible strategies of our character reading practices.

In Section 2, I review the current case in favour of the theory theorist position on character trait attribution, beyond the detail given in Chapter three. I criticise the position on the grounds that, firstly, it seems to only apply to the reasoning that results from the attribution, whereas there is scope for simulation in the attribution itself. Secondly, the theory theorist assumes that the search for simulation is only to be conducted within the replicative processes of practical reasoning—there is no reason in principle why this need be so. In section 3, I make the case for simulative processes being involved in the generation of trait attributions through the reuse of memory, though generally such attributions involve theory alongside simulation. In section 4, I argue that we can use traits in simulations to make predictions of others' behaviour. I give examples of 'enactment imaginations', where I argue that although character information only tweaks inputs to simulative practical reasoning procedures, character reasoning in this way is genuinely simulative when considered holistically. Essentially, when we consider how one can know what sorts of dispositions towards behaviours and psychological states particular characters stereotypically endow, then we can directly reason according to character by simulating the personality of the target.

2. Challenging a Theory Theorist View of Character Reading

Here, I analyse the only discussion of simulation and theory in trait attribution that I could find: the brief discussion in Westra's (2018) account of trait attribution. Accepting that the case for wholly theoretical trait attribution is plausible, in this section I will show that the case against simulation is, whilst intuitive, open to critique.

2.1 Traits and Practical Reasoning

Whether trait attribution involves theory or simulation can be related to its function in reasoning about others, as we do with mental states. The story for mental-state attribution is that 'practical reasoning' allows us to either theorise about (or simulate) the mental states, the beliefs/desires, and the emotions of others, and so reason about why they made certain actions and what actions they might take in the future. 'Practical reasoning' here means reasoning both about one's future actions (what should I do here?) and others' future actions (what will they do here?). It is, essentially, reasoning with a practical conclusion: I will do X/they will do X.

Theoretical practical reasoning is a standard affair. The practical reasoning itself is the process of making theoretical inferences based on one's knowledge of the situation and the target. For example, watching a drunk professional fighter suddenly tense up and loom over someone that was irking them, we might infer that they will start a fight. Such an inference could be made based on our theoretical knowledge about the propensity and capability of the fighter's use of violence in his life, coupled with knowledge that drunkenness lowers inhibitions, et cetera.

Simulation in practical reasoning, on the other hand, is generally thought of as a simulative use of one's own decision-making system. It purportedly functions in the following way: I take my own decision-making system 'offline' for use in lieu of the target's. I create pretend beliefs/desires that the target is thought to possess, where these are fed into my decision-making system to output a pretend (because offline) decision.[71] For example, the belief/desire pair might be the belief that 'there is water in the fridge' and the desire for 'water'. Inputting these into my own decision-making system, I would choose to get the water from the fridge. I then use this output to form a genuine belief about the target: they will make the decision to go and get water from the fridge.[72]

As it pertains to practical reasoning and character traits, Westra claimed that theoretical inference ought to be the default assumption for explanation (though, as I will note in more detail below, what he means by practical reasoning is slightly nebulous). This is because the counterpart of simulation does such a poor job of the task, even when considered as a hybrid account. For example, he noted that "character traits are not the

[71] At least on Goldman's account, it is a belief/desire pair. However, in principle, other mental states such as emotions may be inputted into a decision-making system.
[72] To do this, one must also quarantine one's beliefs and desires that are not shared by the target.

sorts of things that could figure in practical reasoning" (Westra 2018, p. 1223), though "beliefs about one's own character traits could" (ibid, footnote 3). Indeed, "any effect of character on practical reasoning is bound to be oblique: it may affect the kinds of beliefs and desires we form in the first place, the extent to which we deliberate before acting, or the relative importance that we assign to particular desires" (ibid, p. 1224). The upshot for him is that it gives him the space to craft a theory theorist account of how traits translate to behaviour and how they relate to mental states; this is his predictive processing account, which I mentioned in the introductory chapter.

What I will do in 2.1.2 is unpack the objection to simulation in more detail, so that we can be clear on where to challenge it. First, though, a note about the *theoretical* story as it pertains to practical reasoning.

2.1.1 Westra's Account of the Social Cognition of Character

Even though simulation is the focus, it appears that Westra's view is that traits themselves cannot enter into practical reasoning *at all*. It is not clearly stated, but it appears that even if practical reasoning is merely construed as reasoning to a practical conclusion, it requires mental-state attribution. In this sense, Westra is very clearly in the mindreading 'primacy' camp: our understanding of others is achieved primarily by reasoning about their beliefs and desires.[73] I, however, need to endorse the claim that practical reasoning performed theoretically *can* involve traits. This is because of my folk-psychological pluralist commitment outlined in the introductory chapter. Indeed, such is the core of Kristin Andrews' pluralist challenge of the relationship between mindreading and trait attribution—she claimed that one can attribute traits without thereby attributing mental states (Andrews 2008, p. 16).[74] Traits must figure directly in the practical reasoning process when character reading, otherwise character trait attribution is not usefully distinct from mindreading; folk-psychological pluralism would be false.

I take it that one *can* predict or explain someone's behaviour merely according to their traits, without any attributions of mental states being made. It seems to me that an adequate theory of traits should allow one to (situationally) directly infer behaviour from traits, such as predicting that the generous person will tip well in the restaurant. I take it that this is something one *can* do, and that doing so does not require the

[73] See 2.3 in the introductory chapter for the initial discussion of the primacy of mindreading.

[74] I disagree with her on the specifics, to be detailed in Chapter five, but we both agree on this core claim.

attribution of mental states to the target; it is merely of the details in the theory that allow for this that is wanting.[75] That said, there is clearly a relationship between mindreading and trait attribution that warrants explanation, and it is Westra's understanding of this requirement that drives his rejection of pluralism.

Westra takes folk-psychological pluralism, the claim that we socially cognitively reason with things like traits or schemas as well as mental states, to be an alternative to theory-theory or simulation (2018, p. 1224). I think this is strange because Andrews, for example, thought that a theory of traits that relates traits to behaviour can be used to predict the behaviour of others (pp. 16–17).[76] That said, the issue seems to be in the assumption that practical reasoning must take in only belief/desire inputs. This makes a lot of sense for the simulation view discussed below, but I am not sure why it necessarily pertains to theory, given that practical reasoning is merely reasoning to a practical conclusion. The reasoning seems to be that for methods other than belief/desire reasoning, there is an explanatory gap in the explanation of the real relationship between mindreading and trait attribution:

> [The] limitation of the pluralist account of character traits is that it cannot explain the empirical relation between trait attribution and mental-state attribution (…) these two forms of reasoning seem to be causally interrelated, both at the behavioral and neural levels. But on the pluralist account of trait reasoning, mental state information is never involved. This is by design: the pluralist's goal is to show that behavioral prediction and interpretation can happen in the absence of mental-state attribution (Westra 2018, p. 1226).

Firstly, I note that a lack of mental-state attribution in trait attribution does not imply a lack of mental-state information—this is an error that both Westra and Andrews seem to make. I detail this error, in relation to what I call the dependence claim, in Chapter five. I will also show in Chapter five why there *is* such a relation between mindreading and trait attribution, despite an absence of mental-state attribution in some trait attributions. As such, I shelve this discussion of theory and practical reasoning here because *this* chapter deals with my differences with Westra (and any other theory theorists about trait attribution) in relation to *simulation*. The following chapter is related to the content here in that it deals with my treading of the fine line between being

[75] Westra, for example, critiqued how a pluralist account of moving from traits to behaviours might be given (2018, p. 1225). Despite my endorsement of pluralism, I am not convinced that Westra's particular criticisms will apply to my account. This is because Westra's arguments assume no involvement of mental state information, but I think that such is possible, despite a lack of mental-state attribution. I argue for this in Chapter five in my discussion of the 'dependence claim'.

[76] However, this may just be an artefact of my understanding of what theory is (given in Chapter one).

sympathetic to the mindreading primacy view whilst also being a folk-psychological pluralist myself. I needed to make this point here because practical reasoning is under discussion; that said, I return to the task at hand: unpacking the theory theorist's rejection of *simulation* in order to show where it can be challenged.

2.2 Unpacking the Theory Theorist's Rejection of Simulation in Trait-Reasoning

One of the relevant claims is that traits are not the sorts of things that could figure in practical reasoning, so as it pertains to traits, perhaps what is meant is something like the following: I do not think to myself "well, I am a greedy person, so I am going to eat everything in my fridge." Indeed, deciding actions based on one's traits as motivating factors in this way might well be living in Sartrian bad faith; hence, this is not really a proper application of practical reasoning (Sartre 1996, [1943]).[77] Likewise, when engaging in practical reasoning about others, it is possible to conclude that "they are greedy so they will eat everything", but the point is that this would be arrived at through theoretical inferences from theories of greedy people (assuming an adequate theory that details the relationship between traits, situations, and behaviours)—one could not simulate this because then one would be applying the bad-faith reasoning to oneself and projecting it onto others.

This seems right, so we can take the argument to be something like the following: when practical reasoning is simulative, traits are not the correct sorts of inputs to the process. What the process is, is taking one's practical reasoning systems 'offline', running simulations of what others might do, and imputing the resulting decisions to the targets. The correct inputs are beliefs and desires, and so pure simulation of traits is implausible. However, Westra suggested that in response the simulationist would be pushed into a hybrid position, but this will not do either. A hybrid theorist might maintain that, in trait-based reasoning, these beliefs and desires inputted into simulative practical reasoning may be influenced in their selection by the effects of character (determined, presumably, by theories of the functional role of such traits).[78] However, traits are therefore not directly involved in the resulting mindreading simulation—hence the

[77] If one is convinced that this is indeed practical reasoning, albeit poorly done, I note that the phenomena that Westra and I have in mind are the standardly useful instances of practical reasoning that underpin most deliberations, where those are unlike the bad faith example.

[78] It has been suggested to me that I am assuming that Westra appeals to theory for the mechanism by which character traits influence the mental-state inputs to simulative practical reasoning. I am not sure of what other mechanism would be in play here, given that Westra is a theory theorist regarding character reading, and his discussion of the simulationist's possible approaches consists in discussing this method as a "hybrid account" (p. 1224).

'oblique' effect. Because they are not directly involved, "simulation theory [and hybrid theory] is poorly equipped for reasoning about traits" (Westra 2018, p. 1224). For traits in social cognition, then, the most plausible story (so far) is that traits feature as causal variables which affect, according to theories of their functional role, the inputs to belief/desire practical reasoning.

However, even if we accept that traits will influence the sorts of inputs to practical reasoning, having this 'oblique' effect that is claimed, it is not clear why a resulting hybrid theory of this sort (where a theory of traits informs the inputs to a practical reasoning simulation that outputs pretend decisions) should constitute being poorly equipped for reasoning *about* traits (Westra 2018, p. 1224) rather than *with* traits.

2.2.1 Reasoning about vs. Reasoning with Traits

To be clear, the suggestion that I am making is that the argument of the theory theorist concludes that simulation is poorly suited to reasoning *about* the traits of others, i.e., making a trait attribution. There is a further question, though, of whether simulation is suited to practical reasoning about what others will do, by reasoning *with* traits. This needs to be disambiguated because a) I will argue in Section 3 that trait attributions can be made that are based on simulations, and b) I will argue in Section 4 that we can character read by reasoning simulatively *with* traits.

However, even when reasoning with traits in order to predict and explain a target's behaviour, it is not clear in what sense the theory-theorist understands that character information inferring the inputs to simulation procedures would constitute simulation theory (or hybrid positions thereof) being *poorly* equipped for such reasoning. Perhaps the intention is that because the simulations being referenced in such examples would actually be *mindreading* simulations (given belief/desire inputs), then an explanatory strategy of appealing to simulation for the explanation of reasoning with character would be uninspiring. This, of course, boils down to the claim that there is not a reasonable position that holds that character can be simulated.

2.3 Challenging the Theory Theorist's Case for Character Reading

Whilst I agree that a simulation theory understanding of character traits in social cognition is implausible (given the outcome of Chapter three), an argument can be made for a hybrid account. Here, I outline how the theory-theory assumption for character reading can be challenged. The target to be defended is the hybrid account, where character information influences the inputs to offline practical reasoning. I note that if the effect of character in practical reasoning is oblique, this makes a case against trait-based *reasoning* only. Character traits in social cognition are reasoned about as properties of targets, where first one must *attribute* those properties to the targets. Prima facie, a process of simulation (potentially alongside theory) that results in a trait attribution appears implausible, but if such a case could be made, it would undermine the universality of the theory theorist claims about traits in social cognition. Whilst the mindreading-adjacent exploration of Chapter three was unconvincing in making a case for simulation, I do not think that we can reject that view just yet. I give my account of where simulation can feature in trait attribution in Section 3.

Furthermore, we can challenge the theory theorist on the grounds that process simulations might be involved, and character can enter into them, but perhaps not where the theory theorist expects to find them. Consider that the problem, as the theory theorist has it, is that we can simulate the mental states of others by invoking those states in ourselves, and we can then practically reason about what follows, whereas we cannot invoke traits in the same manner. It does not seem like traits can be inputs for the black box of the offline decision-making system.

I agree that this is reasonable; however, I think that it is somewhat missing the wood for the trees. I think that, ultimately, these concerns about trait simulation are well-motivated, but a holistic understanding of what such processes involve will reveal the simulation of traits in reasoning, such that what is *oblique* is actually *fundamental* to the process. A hybrid theorist of trait-reasoning only needs to claim that process simulations are present in some instances of trait-reasoning. The theory theorist has zoned into the traditional simulative story of offline practical reasoning. As such, finding a process simulation in some *other* aspect of character reading is where the challenge to the theory theorist position can be made.

In sum, the existing case against simulation in the social cognition of character traits is reasonable, but not immune from critique. The existing case does not make any claims about the nature of the attribution of a trait itself, merely the reasoning that is involved in the application of trait concepts. In addition, the arguments against simulation in trait reasoning assume that the only place that simulation might be found is in the (rightly rejected) simulation of the decision-making system; traits cannot be inputs to it. As such, there is space (albeit minimal), for a hybrid account of the right kind to exist. I now turn to showing how simulation features in trait attributions themselves.

3. Simulation in Trait Attribution: The Case of Mental Time Travel

As noted in Chapter three, simulation in trait attribution may be present in two kinds of processes: those that are in-principle inaccessible to consciousness, and those that are in-principle accessible to (phenomenal) consciousness. The purpose of making this distinction is that, in building my case, I give examples (of resulting trait attributions) that are phenomenally conscious and merely potentially phenomenally conscious. Therefore, whether we are aware of it or not, simulation turns out to be present in the process of making a trait attribution.

For ease of consideration, I will consider two cases where one actively deliberates: one with a goal of generating a trait attribution, and one without. Crucially, both cases draw upon *knowledge of the target*: it is in accessing and processing this knowledge that I claim simulation is present. If so, then socially cognitive processes can, at least sometimes, use simulations alongside theoretical inferences in order to result in trait attributions: trait attributions can be based on simulations. These cases are those of 'mental time travel' (Tulving 1985; Suddendorf and Corballis 1997): cases where you remember personal events of your past and imagine personal events in your future.

3.1 Accessing Knowledge of Others in Mental Time Travel

In the first case, you are asked: "Remember when your friend so-and-so received bad service at that restaurant last week? Remember how they reacted? What do you think that says about their character?" Here, you are primed to consciously consider trait-relevant behaviour as you form the goal to make a trait attribution and (upon deliberation) report it. Broadly speaking, the attribution is based on what you know about so-and-so.[79]

The second case results in a trait attribution without the conscious goal to do so. Imagine a therapist and their client. The therapist wants their client to recognise the toxic behaviour of the client's family member that ultimately harms their client. The client has intimated that behaviour of this kind, though not conceived of by the client as

[79] One is likely to have already made trait attributions prior to the task if one knows the person, but these need not have been consciously made. Take this example to be a case of the first time they are explicitly considering the notion.

the therapist recognises it, has been occurring much of the client's life (hence various negative character traits can be inferred). The therapist wants their client to come to the realisation themselves, so they ask: "In situation x, how would your family member act?" Suppose that situation x is one where the client is vulnerable and needs support. The client imagines how their family member would react to the situation. They conclude that said family member would engage in some kind of negative behaviour, and perhaps the client would realise that "they would not help me if that happened" or "they would undermine me if that happened". Through the therapist's prodding, the client comes to realise that kind and nurturing people would not act like that; they start to see the family member's toxicity for what it really is, for example, by recognising their narcissistic character.

Both of these cases are examples of making a trait attribution based on what one knows about a target. What is crucial here is the question of the knowledge about the target that is applied in the examples in order to make the judgment: how is that knowledge accessed? I will argue that such knowledge in the questions above is *reconstructed*, rather than *reproduced* from memory, where the reconstructions constitute imaginings (about the scenarios) that involve both theory and simulation. The theory is involved via one's theories of trait concepts and theories of folk psychology, such that the information accessed in memories and presented in imaginings may well be interpreted according to the theory. The simulation occurs in the simulative processes in the act of generating imaginings from memory.

In the following subsections, I first detail how reconstructing the way in which someone will act from memory is itself a simulative process, and then respond to objections that query the use of imagination/imaginings in these scenarios.

3.2 Simulation through the Generation of Imaginings

The simulations present in trait attributions here are in applications of 'constructive episodic memory'. When drawing upon what we know about someone in order to produce a trait attribution in the examples above, the claim is that we take information from memory and use it to generate imaginings about what the targets will do, based on their past behaviours, and presumed psychological states. The physical mechanism of our constructive episodic memory system is *reused* to create these imaginings/simulations. Thus, the creation of an imagining itself is an application of simulation in the process simulation sense defined in Chapter one.

3.2.1 Constructive Episodic Memory

Here, I will unpack 'constructive episodic memory', which will serve to demonstrate how the imaginings in the above scenarios are simulative. There are two types of memory that are relevant to my claims: semantic and episodic (Tulving 1972). Semantic memory is conceived of as our ability to remember general knowledge, free of the context in which we learned it. An example would be remembering a historical date. Episodic memory is the recall of specific events or 'episodes' of one's past, where this might be remembering what I had for breakfast this morning, for example. There are also two relevant theories of memory recall: 'reproductive' and 'constructive' (Cubelli 2010). Reproductive memory is stored in the manner in which it was received. For example, you hear that the Titanic sank in 1912, and when you need this information, you reproduce the exact fact in the manner that you learned it, not necessarily connected to the context in which you learned it. Constructive memory, on the other hand, is stored in fragments and traces, which are then reconstructed from storage and combined with context and beliefs when we want to remember. So, unlike remembering a fact, remembering an event is a piecing-together of the elements of your experience that you had when you experienced the event.[80]

I take the memory recall in the mental time travel examples above to be episodic, in that the examples demonstrate the reflective rerunning of episodes, rather than the recalling of propositions about situations or behaviours. I also take these examples to demonstrate constructive recall, rather than reproductive recall. This is because I take constructive recall generally to be the correct account of episodic memory. In the first story above, one is directly reconstructing a memory and considering someone's traits as they are represented in that memory. In the second case, one is using what one has experienced about the past actions of another to imagine the other's future actions, where doing so results in a trait attribution.

I take constructive recall to be the correct account of episodic memory because if episodic recall were reconstructive rather than reproductive, we would expect to make certain kinds of memory errors that are indicative of a healthy memory system. For example, we would expect errors to occur when subjects are presented with similarly themed words that are different to those which were encountered previously in trials.

[80] In some sense, your experience of the event is rekindled or reconstructed, whereas the recall of a fact has no necessarily associated phenomenology.

Indeed, this was found across many replications. For example, in cases where participants were presented with words such as 'pillow', 'dozy', 'rest', 'blanket', and 'night-time', the participants with typically functioning memories were more likely to recall having seen the word 'sleep' in a subsequent trial, as opposed to correctly recognising that it was a novel word (Deese 1959; Roediger and McDermott 1995; Verfaellie 2002; Budson, Todman and Schacter 2006; Ciaramelli et al. 2006). Additionally, Loftus (1975, 2005) found that eyewitness reports could be altered by phrasing questions differently. People gave higher estimates of driving speed when cars were described as smashing into one another rather than hitting; furthermore, when the term 'smashed' was used, they were likely to falsely report broken glass at the scene.

Another key reason that episodic memory is taken to be constructive (rather than reproductive) is that it is thought that episodic memory being constructive allows us to *imagine the future*. In essence, the construction and piecing together of elements of memory is what allows us to create simulations of what we might think will happen in our personal futures and the futures of others (Schacter and Addis 2007; 2009; Addis et al. 2009). Schacter and Addis claimed that episodic memory serves an additional adaptive function, to "draw on past experiences in a way that allows us to imagine and simulate episodes that might occur in our personal futures" (Schacter and Addis 2007, p. 778).[81] This built on previous work by Tulving, who argued that episodic memory allows for mental time travel, i.e., projecting ourselves into the past or the future. I now move to discuss both the neural and psychological evidence for this claim, showing how such imaginings tie into simulation.

3.2.2 Imagination as Memory Reuse: Neural and Psychological Evidence

If (in the relevant senses) imagination is memory reuse, then we would expect to find evidence for it in a substantial overlap of neural processing when one engages in recalling past episodic events and when one imagines the future. Furthermore, we should find evidence of impaired future imagining in those with impaired episodic memory.[82] Firstly, several studies found evidence of shared neural activity in the bilateral frontopolar cortex when recalling the past or imagining the future (Okuda et al. 2003;

[81] This is not the only function, nor is it the only possible method for recalling the past and imagining the future. The claim is merely that it is one way in which we can, and do, accomplish such tasks.

[82] For evidence that memory systems are important in generating future imaginings generally, for example in how remembering the past biases our reasoning about future events, see Morewedge, Gilbert and Wilson (2005).

Addis, Wong and Schacter 2007; Szpunar, Watson and McDermott 2007). Secondly, O'Connell, Christakou and Chakrabarti (2015), and O'Connell et al. (2018) demonstrated neural overlap in thinking about simulating one's own future choices and taking on others' perspectives in future decision-making.[83]

Impaired episodic memory also affects our imagining of the future. Take, for example, patient 'K.C' (Tulving et al. 1988; Rosenbaum et al. 2005). Due to an injury, K.C. had lost their ability to employ episodic memory. When asked about what was upcoming in their personal future, they were unable to respond, whether the questions were about what they were planning to do next week, tomorrow or even that afternoon—they described their mental state when trying to imagine the future as "blank" (Schacter and Addis 2007, p. 779). A patient with similar issues can be found in Klein, Loftus and Kihlstrom (2002). Furthermore, Hassabis et al. (2007) studied five amnesiacs who had difficulty generating new imaginings that had any detail or coherence to them. In fact, they noted that "the patients' imagined experiences were strikingly deficient in spatial coherence, resulting in their constructions being fragmented and lacking in richness" (p. 1729).[84]

As such, there is much evidence that episodic memory is, in part, responsible for the construction of our imaginations about the future.[85] Indeed, there is a significant neural overlap in processing which is indicative of simulation as per the neural reuse (Hurley 2008) discussed when explicating simulation in previous chapters. To be clear, reusing one's memory to construct an imagining is a simulation constructed from memory systems, as evidenced by the neuro-psychological data above. Therefore, deliberations about what people known to us will do (that result in trait attributions) employ simulations in accessing our knowledge of how those people have acted in the past in order to construct our imaginings of how they will act in hypothetical scenarios. The trait attributions made in the mental time travel cases above, then, are a result of simulative processes; hence, the theory-theory account is false in these cases.

[83] It might be objected that such imaginings of the future are simply 'recastings' of the past with a presumed change in tense, and thus would not count as true imaginings. However, Addis et al. (2009) found that even when 'remembering' and 'future forecasting' were disambiguated in a follow-up study, there remained a significant overlap of processing in particular neural regions for each task.

[84] This suggests that at least one of the functions of episodic memory in imagining one's future is in providing the details of the spatial setting in which a simulation of the future would occur.

[85] Particularly, it is responsible for imagining one's *own* future, as opposed to generic imaginable events (Okuda et al. 2003). This is in line with both the nature of episodic memory, and with perspective-taking (where imaginings of someone else's future are imaginings of *one's own* future as if one were them).

3.3 Objection: Why is Imagination Involved?

All of the above, of course, relies on there being imagination involved in the cases of mental time travel. The critic might note that now I have made the notions of relevant memory clear, the theory theorist about trait attribution can simply deny that imagination is involved. Why claim that either recalling the past or predicting future behaviour requires imagination? The theory theorist might simply claim that trait attributions made by recalling people's behaviour in the past are just reproductive semantic memory. In essence, on the assumption that a trait attribution has been made previously, they can simply reproduce the attribution of laziness, for example, much like someone can remember a date. Likewise with the prediction of future behaviour, one might claim that one does not need to imagine how someone will act in the future because one can directly predict future behaviour by reference to past behaviour. As such, the theory theorist might be able to derail my argument for simulation by denying imagination from the start.

Regarding this kind of objection, I start by noting that it may well hold for a number of examples that we can suppose in everyday life. This highlights the plurality of folk-psychological strategies that we can employ; I am happy to concede this. It is not, however, an appropriate option for the examples given above. I take these cases to involve imagination because they are episodic in nature—these examples are about memories of events in one's personal history (where in the second case one uses one's memories of events in one's personal history to construct imaginings of the future). As such, the memories are not themselves semantic memories.[86]

Of course, the theory-theorist can concede this point but still insist that imagination is not importantly connected to these examples or crucially, simulation. The claim might be that imagination is involved, but it is essentially epiphenomenal. In essence, one might reconstrue the mental time travel cases above to be memory recall that involves imagination, followed by then *theoretical inference* based on the production of the imagining. As such, the objection goes that even if it does not feel like it, the work is

[86] One might suggest that, in the therapist example, the client may be able to use brute facts about past behaviour in order to come to the trait attribution, as the behaviour merely needs to be known in order to generate future imaginings. However, this example is constructed specifically to ward off this consideration due to the fact that the therapist prods the client to engage in pensive reflections that result in the trait attribution. It is unlikely that the experience of such ruminations on the past actions of others (that are directly pertinent to you) are mere recollections of fact, as opposed to an emotionally weighted reliving of the past.

being done by theory. However, here is the catch: if one concedes, as one should, that memory and imagination are involved in the examples of trait attribution based on mental time travel above, then simulation is *already* involved. Episodic memory *is* what allows one to imagine events in one's future—it is a *reuse* of memory architecture; hence, it is simulative.

In summary, this section has argued that there are cases in which trait attributions can be made that are the result (at least in part) of process simulations, contrary to the theory-theorist position outlined in section 2. I argued that in trait attributions that involve mental time travel, generating an imagining in order to make the trait attribution is itself simulative, given that 'imagining' in such cases is a simulative reuse of episodic memory architecture. I now move to the second part of my argument that simulations are involved in character reading: I will argue that the imaginings are themselves simulations when they are holistically considered as 'enactment imaginations', which are paradigmatic instances of simulation on Goldman's (2006) account.

4. Character Reading

In this section, I develop the claim that one can use traits in simulations that result in, for example, predictions of behaviour. I take it that if such cases, such enactment imaginations (henceforth e-imaginations), do use traits in simulations, then character reading can involve simulation (alongside theory). Goldman noted that "When I imagine feeling elated, I do not merely suppose that I am elated; rather, I enact, or try to enact, elation itself" (Goldman 2006, p. 47). E-imagination for mental-state attribution works in exactly the same way as the simulative story of practical reasoning given in 2.1—it just highlights that there is this distinctive kind of attempted enaction in this kind of simulation.

The case study I will discuss is intended to pump the intuition that a personality has been simulated, and then I will argue that even though mental states might be involved in the practical reasoning, this does not take away either from the distinctive simulative character of the reasoning or, crucially, that traits are fundamentally involved in such a way that the moniker 'character reading' is appropriate. I am not claiming that this could not be achieved solely with theory, but this example is intended to illustrate that, for some people, running a simulation like this may be more helpful than merely theorising.

The examples below are two snippets from the popular New York-based police sitcom 'Brooklyn 99'.[87] In this example, I want to contrast what appears to me to be an instance of the prediction of behaviour based on traits that is a wholly theoretical inference, with one that involves simulation. Hopefully, this should pump intuitions that there is an important difference between the two that understanding the latter as using traits in simulative practical reasoning can capture.

4.1 Jake Peralta Advances the Plot with Character Reading

In the episode 'Chasing Amy', protagonist Amy Santiago is about to take the police sergeant's exam. She is anxious and ends up fleeing a practise test that her partner Jake Peralta set for her. Jake asks the staff in his office: "If you were Amy, where would you be right now?" His colleague, Gina, quips: "Boring pantsuit store, a crossword factory, a museum of retainers and headgear. Is it possible to enter the colour beige?" This seems to me to be a prediction that, whilst obviously unserious and derogatory, is made based on theorising about Amy's supposedly boring character. This is because, in order to obtain this result, all one needs to do is think about boring things in general (beige) and boring things of a particular kind that Amy would enjoy (a crossword factory). Contrast this with the following example of the next relevant scene. Trying to work out where she has fled to, Jake takes his friend and colleague, Rosa, out onto the New York streets in order to try and find Amy. He decides to "try and see the world through her eyes". The following is a transcript:

> Jake: Just know, everything I'm about to say, I say with love. Ok, I'm outside, it's 78 degrees and yet somehow, I'm still cold. Better walk on the sunny side of the street. Wait, did I remember to put on sunscreen? Pause to think, pause to think, yes, I did put on sunscreen, and I bragged about it…all morning. Amy's on the move…and I'm walking, I'm walking (…) I'm walking, I'm walking, and I'm seeing a paper supply store! I wonder if they have those new summer folders.
>
> Rosa: Oh, you're good. I've definitely heard her talk about those folders.
>
> Jake: That's because they have such a fun colour scheme, perfect for vacation organisation.

[87] Obviously, this is a fictional example, but nothing about the scenario or actions discussed are *ceteris paribus* impossible or implausible.

> Rosa: This is chilling.
>
> Jake: Darn it to heck, the store is closed. My bad day just got a whole heck of a lot worse. Time to go…smoke in secret [puffing noises]. And, as I'm shame-smoking, I'm seeing [looks around whilst imitating smoking] …
>
> Jake and Rosa: The library!

This example seems very different from Gina's outburst. Here, Jake is using what he knows about the type of person his partner is in order to simulate how she would act. It seems to me that Jake cannot fail to simulate Amy's personality here, because this is exactly what simulating personality *consists in*: taking their perspective, seeing the world as they see it, and inferring—based on who they are and what dispositions they have—what they would do. As a slightly meta point to this (and to ruin the joke with philosophical reflection), the writers' intention in this scene is clearly to convey that Jake is simulating Amy's personality, where the writers are the ones who know Amy the best. As such, Jake taking on Amy's personality is achieved technically perfectly, hence the comedy in the re-enactment of her particular idiosyncrasies.

This is the example given to pump intuitions that an instance of simulative character reading has occurred. I now show how this relates to both the accounts of practical reasoning discussed, and the claims that character's effect on practical reasoning will only ever be oblique.

4.2 Simulating Character

The example was given to pump the intuitions that despite the technicalities of beliefs and desires being inputted into simulative practical reasoning, there is still something to be said for the claim that character is simulated. My claim is that the whole process that Jake engages in, considered holistically, is itself a simulation of character. If you feel the pull of this intuition, great, but I am not expecting to be convincing merely with that claim. What would be a good argument for this claim? Well, I claimed earlier that the theory theorist was assuming that simulation could only exist within the practical reasoning process itself. The task, then, would be to show how simulation can be evidenced somewhere else. This can be accomplished by showing that the definition for process simulation has been met in the sorts of holistic considerations that I am making. I established in Chapter one that process simulation is the focus of this dissertation; I also gave Goldman's definition of process simulation and showed why it was apt. As a reminder, here is the definition once more:

> Process P simulates process P' = df.
>
> (1) P duplicates, replicates, or resembles P' in some significant respects (significant relative to the purposes or function of the task), and
>
> (2) in its (significant) duplication of P', P fulfills one of its purposes or functions (Goldman 2006, pp. 37–38).

This technical definition (which the astute reader may have noticed has actually served no real purpose up until now, besides being technical for technicality's sake) provides us with a clear set of desiderata for the assessment of the appearance of process simulation in character reading. I will argue, then, that this definition of process simulation is met in cases like the Brooklyn 99 example.

Based on this definition, we need to know four things regarding character reading:

1. What the original process is.

2. What the purpose or function of the original process to be simulated is.

3. In what ways the simulation replicates or resembles this process.

4. How the simulative process fulfils its own function.

For number one, the original process P' is simply living one's life. It is the enaction of one's life in the day-to-day. Here, like simulation, process is used holistically to refer to the sum of all of the processes that are involved in living one's life. In the same way that a process can refer to the particulars of cognition, a process can also refer to wider concepts like living, or working, or undertaking a PhD, for example.

For number two, the purpose of the original process is what the purpose for that person's life is. I make no claims about human nature or our goals here—I take any appropriate formulation of a person's purpose to be acceptable. What matters is that it is an *authentic* purpose, that is, the purpose stems from their genuine self—character reading (when simulative) attempts to understand others by who they are, not by who they appear to be. As such, one and two combine into the target: someone in the process of living, in truly being who they are.

For number three, the way in which the simulation resembles the original process is that in enacting someone else's character, you act and think like them to the extent that you are becoming them—the enacted decision-making of the target is, at its core, an enaction of *who they are.* The simulated decision-making of the target needs to be formed and constrained by who they are as people, otherwise the simulation of purely rational decision-making would always make the same predictions of behaviour no matter who the target is this is clearly not the case. Knowledge of someone's past behaviours helps in becoming another, but for these to be relevant to the decision-making at hand, these are going to need to be the past histories of disposition manifestation, that is, histories of their character. In the Brooklyn 99 example, Jake's replication of Amy's personality is carried out so well that he successfully replicates the way in which Amy lived her life such that he finds her before it is too late for her to take the real sergeant's exam. The key point of consideration is that, when zoomed in on the underlying processes of practical reasoning, this can be described as character information tweaking the inputs to practical reasoning but considered holistically it is part of the simulation of character.

For number four, the simulation of character fulfils its own socially cognitive purpose, in that it enacts Amy's character to the extent that Jake and Rosa can work out where to go in order to find Amy—this is essentially a trait prediction.

Given, then, that we can meet the four conditions for the technical definition of process simulation when we consider character reading holistically, we can tentatively conclude that the 'where' of simulation in character reading need not be confined to replications of decision-making systems in practical reasoning, as the theory theorist holds.

In sum, the claim that traits do not enter into practical reasoning themselves, and that traits only feature obliquely as causal variables determining the beliefs and desires of targets to be simulated in practical reasoning simulations, obscures the fact that the dispositional component of character *just is* a disposition towards certain behaviours and psychological states. The reasoning about this would function identically to traits acting as underlying causal variables that determine the beliefs and desires of the targets to be simulated. However, this does not mean that this is only an oblique effect, as opposed to being fundamental to the process, and as opposed to being recognised holistically as the simulation of another's character. Beginning with the intuition-pumping case regarding Jake Peralta and his fiancée Amy, I fleshed this claim out by noting how a holistic consideration of character reading meets the definitional markers of process simulations on Goldman's account. As such, we ought to understand that traits can indeed enter into simulations alongside theorising; as such, the social cognition of character ought not to be conceived of as necessarily theory-theory in nature. Having given some reasons for understanding the entire process that Jake goes through as a character simulation, there remains one final quick objection that I ought to address (seeing as I first raised it): the combination problem, which I first raised in Chapter three.

4.3 Is my Account of Character Reading Subject to the Combination Problem?

One might think that my account is subject to the combination problem. After all, I noted of the 'generate-and-test' and 'mirror resonance' methods that they were subject to a combination problem in the sense that they could not explain how trait-relevant markers combined into trait attributions.

To this I respond that, firstly, the trait attribution has already been made on my account of character reading—there is the representation of the person as having particular traits that are simulated, and so character reading relies on your having already attributed a trait. Secondly, the combination problem does not arise for my account of simulation in trait *attribution* because theory is involved. As I noted as a solution to the combination problem, a theory of traits may well be able to provide the sorts of information required to move from behaviour to traits—my account is a hybrid account precisely because I think that, in instances where trait attribution is simulative, theory necessarily serves this role.

However, the critic is right to pick out the fact that there is still a small missing puzzle piece here, namely a precise formulation of the theory of traits that allows one to move from behaviours to traits (given that one's behaviours tend to underdetermine one's traits). That said, I have noted that a solution is not in-principle impossible because trait predictions and explanations without mental-state attributions *are possible*. My account of theory in Chapter one provides some constraints on the sorts of information that might be added to the theory, and I detailed exactly what character traits are in Chapter two in order to aid in such a task, so I have contributed towards such an endeavour. Furthermore, in Chapter six, I will discuss an account of how we might move from seeing facial displays and their requisite emotions to then attributing traits, so part of the story is being told there even if I do not have all the answers here. I see the main contribution of this chapter as situating the place of *simulation* in character reading, given that theory's role has always been presumed by scholars working on these issues.

Conclusion

In this chapter, I argued that despite the plausible reasons for being a theory-theorist regarding character traits in social cognition, there is a good case to be made for the presence of simulation. I argued that simulations can occur in trait attributions when one generates an imaginative simulation of a target that reuses episodic memory architecture, and also that—in making predictions and explanations of behaviour—character traits themselves can feature directly in a simulative process. This is because a holistic understanding of character information tweaking the inputs to practical reasoning is simulating one's dispositions towards certain behaviours and psychological states itself; this is a simulation of personality. As such, this chapter has given my account of character trait attribution and character reasoning and argued for a hybrid account; simulation may not be common, but it *is* there in the cases I have discussed. Now, my discussion of character reading has thus far, particularly here and in Chapter three, assumed some kind of connection or relationship between character reading and mindreading. I move, in the following chapter, to detail this relationship.

Mindreading and Character Reading: Conceptual and Explanatory Dependence

1. Introduction

We have seen how character reading is appropriately categorised by theoretical and simulative accounts of its functioning, though in a way that differs from mindreading. Until now, I have assumed both a distinction and some kind of implicit relationship between the two folk-psychological skills; now is the time to detail this. Whilst Chapter six, in-part, argues for the ontological dependence of character reading on mindreading through character reading's dependence on the physical mindreading mechanism, this chapter details the conceptual and explanatory dependence of character reading on mindreading. The dependence relations I detail here demonstrate that the current primacy of mindreading (as our main folk-psychological skill) is not undermined by pluralist folk psychology, at least for character reading, but nonetheless one can still maintain an otherwise pluralist approach to folk psychology.

In Section 1, I reintroduce the pluralist challenge of mindreading's primacy, and focus on Kristin Andrews' pre-emptive response to a potential defence by primacists. This is in her denial of a 'shorthand' claim, i.e., the denial of the view that character trait attribution is necessarily shorthand for mental-state attribution. Whilst sympathetic with Andrews on this point, I disambiguate her presentation of the shorthand claim from a 'dependence' claim that she also makes. Section 2 defends this disambiguation by unpacking the shorthand and dependence relations, in order to demonstrate that whilst shorthandedness entails dependence, dependence does not entail shorthandedness. Section 3 argues for a dependence relation between trait attribution and mental-state attribution, as I argue that trait attributions depend on one's self-knowledge of one's own mental states. Section 4 argues that on Andrews' account of folk-psychological explanation, there is no good reason for why trait explanations are satisfactorily explanatory—I plug the gap in this account by demonstrating how a dependence relation between character reading and mindreading can make trait attributions explanatory.

2. Mindreading Primacy and the Pluralist Challenge

2.1 Mindreading Primacy

One of the key pluralist claims is that mindreading is not the main method of cognising socially—one can also (and often does) attribute character traits, adhere to social schemas and scripts, shape the contents of others' minds in a regulatory fashion, and attribute stereotypes, et cetera. It is the case study of character trait attribution, as a distinct folk-psychological skill from mindreading, that Kristin Andrews uses to argue that mindreading is not the prime method by which we understand others (Andrews 2008, 2012). Her main claim is that character trait attribution cannot be identical to mental-state attribution because one can attribute character traits without attributing beliefs and desires; thus, if mindreading is not necessarily involved in other folk-psychological skills like character reading, then the primacy of mindreading is undermined.

That mindreading is not identical to character reading seems obvious, but there might still be room to retain mindreading's primacy if mindreading were somehow involved in or underpinned other folk-psychological skills. Hence, it is in the assessment of a potential objection from her opponents' on this score that I think a key relation is implicitly discussed and conflated with Andrews' main claim; uncovering this will help us understand the true relation between character reading and mindreading, i.e., one that retains the primacy of mindreading but captures the spirit of pluralist folk psychology. First, I motivate and explain her framing of this objection.

2.2 Disambiguating the Dependence Claim from the Shorthand Claim

As Andrews notes, those who are sceptical of a pluralist approach that denies the primacy of mindreading might propose a 'shorthand' relation between character reading and mindreading (Andrews 2008, p. 19). If character reading is shorthand for mindreading, then not only is the primacy of mindreading retained, but character reading is merely a particular description of mindreading processes. As such, the paradigmatic case study for pluralistic folk psychology would be undermined. Here, I quote Andrews from her 2012 book, in which she describes and contextualises the shorthand claim:

> For example, personality traits might be associated not only with behaviors but also with intentions, beliefs, desires, and other internal contentful states (…) If this view were right, then a trait attribution would be shorthand for the attribution of some set of beliefs, desires, and so forth (…) If trait attribution is simply shorthand for belief or desire attribution, then our ability to predict behavior by attributing personality traits relies on a prior ability to attribute beliefs and desires (…) Understanding traits in this fashion reconciles at least

> part of the social psychological picture with the traditional philosophical one, and thus the evidence from social psychology about trait attribution would not undermine the claim that we predict behavior through the attribution of beliefs and desires (Andrews 2012, pp. 100–101).[88]

Andrews gives reasons for rejecting the shorthand claim, but these are not the focus here (Andrews 2012, pp. 101–105). The focus is that there are two claims being made in the quotation above that should not be conflated: a 'shorthand' claim, and a 'dependence' claim. Whilst Andrews builds a case against the shorthand claim, I take the dependence claim to be a key and unrivalled explanation for the relationship between character reading and mindreading. The *prima facie* distinction is thus: The shorthand claim is that when we make a trait attribution, we are attributing some specific mental state content. Therefore, the trait attribution is 'shorthand' for mental content attribution. For example, attributing generosity *means* attributing, for example, beliefs and desires that are trait relevant, like desires to share. This is a claim that both Andrews and I reject.[89]

The dependence claim, by comparison, is such that trait attributions are dependent upon attributions of mental states to others. Dependence is a relation that will be unpacked below, but *prima facie* the dependence claim deserves to be distinguished from the shorthand claim because it is possible that one could attribute a trait based on an understanding of traits as mentalistic concepts but not specifically attribute mental content like belief/desire pairs. To understand someone as 'generous' may not require occurrent belief/desire attribution, but making the attribution might require a prior understanding that the person can have mental content of a kind associated with the relevant trait; this chapter argues for this claim. In this sense, if the shorthand claim is true, then trait attribution is dependent on mental-state attribution; it does not follow from this that if trait attribution is dependent on mental-state attribution, then the shorthand claim is true. Likewise, of course, mine and Andrews' denial of the shorthand claim does not entail that there is no such dependence relation.

[88] Andrews also briefly mentions the shorthand claim in her 2008 paper, p. 19.

[89] My reasons for rejecting the shorthand claim are different from Andrews'—to me, it seems clear that one can consciously attribute traits without also attributing particular (or sets of) mental states; in this sense, the shorthand claim is something like the implausible 'identity claim' of character reading and mindreading in disguise. Secondly, it does not seem, to me, to adequately capture the meaning of a trait attribution by claiming that it merely means an attribution of some mental states—what is meant by a trait attribution is that an individual has a particular character of the kind detailed in chapter two, not just mental states that are indicative of said character.

Thus far, we can see that the pluralist stance—at least that which is given in Andrews' seminal work—is to claim that there is no distinctive relationship between character reading and mindreading except in so far as they are both folk-psychological skills; the shorthand relation *would* have reduced trait attribution to mindreading, but it is false. My approach in this dissertation has been to be sympathetic to the pluralist approach, but it is here that my account differs. My account, focussing on the dependence relation, will establish that character reading and mindreading have a relationship predicated on dependence, in which one may still accept the primacy of mindreading without denying the distinctiveness and relative commonality of the application of the folk-psychological skill of character reading. In order to do this, it is imperative to be much clearer on what is meant by both the shorthand and dependence relations.[90]

3. Unpacking the shorthand and dependence relations

Positions endorsing shorthandedness or dependence both claim that character trait attributions are associated with mental content attributions, but it is the nature of the association that differs. Shorthandedness requires mental content attribution when trait attributions are made, whereas the dependence relation is not constituted by this. What kind of relations are these anyway, and what does it mean to *depend* on something?

3.1 The shorthand relation

I am taking the shorthand relation to be psychological in kind. Given this, might shorthand mean 'analytically equivalent in virtue of meaning', or something like 'a disposition to defuse the trait explanation in favour of a mental state explanation when asked'?[91] Shorthand attributions are not dispositions to give particular explanations—Andrews intends that shorthandedness should refer to something occurring during the attribution, not after the fact. If two propositions are analytically equivalent in virtue of meaning, then if one proposition is true, the other must be true and *vice versa*. Of course, if two propositions express the same meaning, then they are the same proposition; the equivalence in 'analytically equivalent' is in relation to the way in which the propositions are expressed (for example in English or in logic). Hence, for trait attribution, to attribute the trait of generosity to someone is to be committed to the claim that you

[90] My move from talk about a shorthand/dependence *claim* to a shorthand/dependence *relation* is merely a move from the analysis of the fact that that a claim has been made about the existence of such relations, to the analysis of those relations themselves.
[91] I owe this clarification to an anonymous reviewer of a paper that discussed this topic.

have attributed to them (for example) beliefs and desires associated with generosity, on the grounds that claiming that someone is generous *means* the same thing as claiming that they have certain beliefs. Likewise, if they could not be attributed the beliefs, they could not be attributed the trait. In this sense, the shorthand claim (as a psychological relation) is about meaning, about what the concepts involved in the attributions refer to. With this clarity, it is apparent why the shorthand claim is false: whilst one could clearly *mean* to attribute certain beliefs to someone when making a trait attribution, one need not do so, and they may even intend not to.

3.2 The Dependence Relation

My strategy for the rest of Section 3 is to clarify some key aspects of what a dependence relation entails, to set up the notions of dependence that I evidence in subsequent sections. Because character reading and mindreading are more than merely the attributions that Andrews discusses (recall my definitions of these terms from Chapter one), my case for the truth of the dependence claim is split into two arguments: Section 3 will argue for a way in which character trait *attribution* depends on mental-state attribution; Section 4 argues for character *reasoning* depending on mindreading through an argument that trait-based explanations of behaviour depend on mental state explanations of behaviour.

3.2.1 A Constituency Relation of Contents

The dependence claim might be such that mindreading bears some kind of *constitutive* relation to character reading, where elements of mindreading constitute elements of character reading. Alternatively, it may be that there is some kind of *contingency* relation between the two, such that whilst one does not constitute the other, it is in some sense contingent on its functioning in order to function itself.[92] It is not clear to me what opting for contingency gets us, as it seems to 'kick the can down the road' regarding the provision of an explanation of the relevant relation—on this view, we would still be required to produce some account of functional contingency that does not appeal to shared parts in its functioning. The constituency claim, meanwhile, can explain contingency in functioning easily by appeal to shared elements that constitute both abilities. Furthermore, this seems to fit better with the account I gave in Chapter four,

[92] This distinction was suggested by an anonymous reviewer of a paper that discussed this topic.

where elements of mindreading are used in character reading when becoming another for socially cognitive purposes.

So, assume a constitutive relation. Is this relation one where the *capacity* for character reading is constituted by a capacity for mindreading, or where some *content* of character reading is constituted by the content of mindreading? I claim *both*, but I focus on the latter in this chapter, and in so doing demonstrate how the former is intelligible. A case for the *capacity* for trait attribution constitutively depending on the capacity for mindreading will be given in Chapter six. As it stands, I claim a constitutive dependence relation between character reading and mindreading—I cash this out in the following section by suggesting a conceptual dependence relation between the two.

3.2.2 A Dependence on Mental Content Concepts

Unlike the content of the shorthand claim, I do not hold that mental content concepts are necessarily attributed in trait explanations. As such, how might an association between trait attributions and mental-state attributions which, crucially, maintains the primacy of mindreading be cashed out? Two potential options are that trait attributions could attribute mental content concepts *generically* or *dispositionally*.[93]

How satisfying are these options? Firstly, it is clear that they can be distinguished further. The term 'generically' implies that one would not be attributing specific mental states; rather, it implies the attribution of mental states generally, such that 'this thing is a possessor of mental states'. From this we can distinguish further among senses of 'generic', because an equivalent way of noting that people have mental states is to attribute to them some non-empty set of mental states. This attribution could be *truly* generic, in essence an attribution of a non-empty set of *any* set of mental states, or it could be *relevantly* generic, such as the non-empty set of some set of mental states which possessors of trait X are likely to manifest. On the other hand, the term 'dispositionally' implies that one would be attributing a disposition to possess or otherwise manifest a mental state of a certain kind. As such, in theory one could distinguish between an attribution of a disposition to manifest a *specific* mental state, a disposition to manifest mental states *generally*, or a disposition to manifest some of the contents of a set of mental states that are *relevant* to possessors of the trait.

[93] I owe this distinction to an anonymous reviewer of a paper that discussed this topic.

Truly and relevantly *generic* attribution are both unsatisfactory for the dependence relation. This is because generic attributions are still attributions; hence, they are restatements of the shorthand claim. The distinction between kinds of genericness is important though, because I am committed to a dispositional sense, where the obvious distinction to make would involve claiming that a trait attribution is associated with mental-state attribution because it involves an attribution of a disposition to manifest some of the contents of some set of relevant mental states as being related to the attributed trait—the *relevant* case, rather than the *truly generic* or *specific* sense.

If trait attributions attribute mental content concepts dispositionally, then trait attributions depend on mental content *concepts*. This is because it is through the possession of mental content concepts that it is intelligible for one to attribute a disposition to manifest relevant mental states for a given trait. If trait attributions require disposition attributions of this kind, then trait concepts *themselves* must depend on mental content concepts—I argue for this in Section 3. Finally, if the concepts we use in a trait attribution are dependent on mental content concepts, then our capacity to attribute traits must also depend on mental content concepts—we can be satisfied that this constitutes the kind of dependence claim that I am interested in.

Before I proceed to a summary, a quick pre-emptive note: one might suppose that a conceptual dependence between trait concepts and mental state concepts is, strictly speaking, a different issue from dependence between trait attributions and mental-state attributions. If so, a dependence relation between concepts might obtain, but might not be sufficient to establish, a dependence relation between character reading and mindreading. In response, you might ask: Can you attribute concepts that you do not have? I take it that you cannot, because you do not possess them to attribute them; any attribution made would be an attribution of some different, or (empty) placeholder, concept. Hence, I take this to give me license to talk of concept possession as applying to concept attribution in these cases. Of course, it might be so that an attribution mechanism for traits need not depend upon an attribution mechanism for mental states, but I show in the following chapter that the mechanisms that we have do so.

It should also be noted that later in this chapter I discuss *explanatory* dependence, and in Chapter six, *ontological* dependence. 'Explanatory dependence' refers to explanations of behaviour being explanatory because of a dependence relation between character reading and mindreading. In this sense, explanatory dependence is a more general form of dependence on reasoning that bottoms out in the conceptual dependence detailed here. A conceptual dependence is, of course, a type of ontological dependence; however, the ontological dependence I discuss later is specific to the context of the ontology of the physical construction of the cognitive mechanisms for character reading and mindreading—I call it 'ontological dependence' simply to distinguish the other relevant sense in which character reading depends on mindreading.

In summary, this section has distinguished the shorthand relation from the dependence relation, and unpacked the dependence relation. The dependence relation is about how dispositional attributions of mental states, made when attributing traits, are constitutively related to the trait concepts attributed (where the trait concepts are the relevant *contents* of character reading that warrant explanation by the constituency relation). If this dependence relation obtains, then the primacy of mindreading is not undermined; this is due to the reliance on the contents of mindreading. This does not entail, however, that folk-psychological skills such as character reading should not be considered under a pluralistic lens, given that character reading is not shorthand for mindreading.

Now, I need to provide some evidence that the relation does actually obtain in the way that I have outlined. To that end, the rest of the chapter makes this case. In the following section, I will argue that trait attributions require self-knowledge of one's own mental states; hence, the conceptual content of character reading constitutively depends on the content of mindreading. I will then defend this claim against an objection that there is evidence that some autistic individuals can make trait attributions without such knowledge of mental state concepts. In the final section, I will argue that what makes trait explanations explanatory is better captured by accepting the dependence relation, as opposed to Andrews' account.

4. Trait Attributions Depend on Self-Knowledge of our own Mental States

Regarding the many conditions that one needs to meet in order to be said to possess, to have, a trait concept, one of them is the knowledge that traits are the sorts of things that you yourself might have. If I have a concept, I can use it. If I can use a concept, I can use the concept to refer to its extension, and so I can make trait attributions. When it comes to the character trait concept, at least one of the necessary possession conditions is that you must have the knowledge that traits are things that persons, including yourself, can have.[94] If you do not know that a trait concept's extension is some property that you could have under some circumstances, then, frankly, you do not understand what a character trait is.

That traits are the sorts of things we can have derives, at least in part, from one's knowledge of one's own current mental states. Here, 'self-knowledge of one's own current mental states' refers to that kind of knowledge about our own mental states constituted by our mental-state attributions to ourselves. It is in this sense that we can monitor (merely keep track of) our mental states. I make no claims about the overall accuracy of these attributions; I assume at least enough accuracy for having such an ability to be useful. Simply enough, if I am able to think of myself as happy when I am happy, this constitutes an example of the self-knowledge of mental states. My claim is that when making a trait attribution, I require the ability to attribute mental states to myself, because such a process of trait attribution requires the knowledge of my mental states (and others') in order to pick out the trait-relevant psychological properties—'Am I really a cantankerous person? Do I often think and act that way?' This, of course, need not be a conscious process. Regardless of this, if trait attribution requires self-knowledge of mental states, then trait attribution depends on the ability to attribute mental states.

4.1 Character Traits, as Theoretical Concepts, Require Self-Knowledge of Mental States

My argument for the claim that trait attribution requires self-knowledge of mental states is fleshed out by arguing that the way in which character trait concepts are acquired requires self-knowledge of mental states. In order to see more clearly why trait attribution requires self-knowledge of mental states, let us think more broadly about how someone may come to know about character traits. Particularly, as well as knowing that traits are the sorts of things that you can have, they are also the sorts of things that *others* can have. If you do not know either that character traits are things that people other than yourself can possess, or that they are things you can possess, then you do not really possess the concept of a character trait. To know both of these things, you may

[94] This is not to say that the concept can be completely determined by necessary and sufficient possession conditions; there are good reasons for thinking that few, if any, concepts can be determined in this way (Margolis and Laurence 1999, pp. 13–27). I only claim here that there exists this necessary condition, and that knowing that one can have character traits is necessary to have the concept. However, I will later also claim the existence of a complementary necessary condition—that of the knowledge that traits are things that other people can have.

adopt one of two views about the kind of concept acquisition that a character trait concept requires.

4.1.1 Observational Concepts

The first view would be to suppose that this knowledge stems from the *character trait* concept being an 'observational' concept (Peacocke 1992). That is, we can understand the concept primarily from our own experience. Borrowing an example from Peacocke (2008), I can know what *pain* is because it is that which is the same property that *I* have when I am in pain (and thus I recognise it in others). Peacocke noted that we would initially apply some unstructured recognitional concept, one that:

> picks out the property [of pain] not by some further concept or description, but rather by the fact that pain is the property of the sensation to which the thinker is rationally and causally sensitive in his application of the unstructured concept. It is the property to which a thinker is rationally responding when he exercises a recognitional capacity for his own pains. (2008, p.172)

On the observational view, then, traits would be known from our own experiences, and we could then recognise them in others because we recognise them in ourselves.

4.1.2 Theoretical Concepts

Not all concepts can be acquired as observational concepts are. The *quantum computing* concept cannot be known from our own experiences—the concept is possessed by way of understanding a certain theory. Call this the 'theoretical' concept. On the theoretical concept view, traits would be known by grasping a certain theory about what traits are, and thereby granting knowledge that the concept may apply to ourselves as well as to others.[95]

I think it unlikely that traits are observational concepts, given that there does not seem to be the same sort of privileged access to those properties of experience that denote traits, compared to, say, experiencing pain. By this I mean that it is not obvious if there is any particular property that we can attend to when introspecting on our traits, or any particular property to recognise in others, in the same way that one can do for pain. Employing a recognitional capacity for our own traits appears to be—as Judith Jarvis Thompson might have said (as she did of the property of 'goodness')—"metaphysically dark": it is really not clear what we might be attending to and what its nature is (Thomson 2008, p. 11). Nevertheless, if traits are observational in kind, then one would need self-

[95] Recall the discussion on theory in chapter one for how this might function.

knowledge of one's own mental states in order to perform this recognition that traits are things we might have.

On the other hand, traits as theoretical concepts appear to match our experiences better, not to mention accounting for much of the presence of theoretical inference in character reading.[96] We possess certain theories of particular character traits, and of the idea of traits themselves, as tendencies towards particular psychological contents and actions (as detailed in Chapter two). Now, with traits being theoretical concepts which are possessed by possessing a theory of them, we can note that one needs to reflect *at least* on the behaviour of oneself and others (or at least apply knowledge of these sorts of behaviours) in order to recognise a trait. Even if we are only thinking about behaviour (of ourselves or others) in order to pick out trait-relevant properties, we need to use our knowledge of our own mental states to do this because, on the theoretical view, the fact that others might act like you is grasped, at least in part, by understanding that others might have similar reasons for acting. Particularly, knowing that you are a thing that has thoughts and feelings, that you often act as a result of your thoughts and feelings, and recognising that others might act like you includes understanding that they will be often motivated to act by their own thoughts and feelings. The *relationship* between thought and action is something grasped by the knowledge of our own mental states, where this knowledge is required in order to make trait attributions of others.

As such, that trait attribution depends on mental-state attribution is evidenced by the fact that however trait concepts are acquired (though it is most likely theoretically), one requires a pre-requisite self-knowledge of mental states—this cashes out the claim made on page five that trait concepts themselves depend on mental content concepts, because the contents of the self-knowledge of one's own mental states are mental content concepts.[97] Having argued for the evidence of the dependence relation on the grounds of how we acquire trait concepts, I will now pose and respond to an empirical objection.

[96] Indeed, there are parallels between observational accounts of concepts and simulative accounts of mindreading/character reading. In principle, perhaps some complex concepts are hybrid concepts acquired with both observation and theory; nevertheless, my point stands that theory is involved.

[97] Perhaps this is where some confusion about trait attribution stems from—recognising trait-relevant properties may only involve a comparison between types of behaviours (and not mental states); hence, one denies the shorthand claim. However, in order to recognise the trait, one needs to have some ability to monitor one's own current mental states, which are evaluating these behaviours.

4.2 Objection: Some Autistic People Attribute Traits Without the Ability to Attribute Mental States

We might think that my argument only holds in so far as there is no empirical evidence that we do, in fact, attribute traits without an ability to attribute mental states. In defence of her general pluralist stance that character reading and mindreading come apart, in her 2008 paper and 2012 book Andrews claims to present some such evidence: she appeals to an 'intervention method' for helping autistic children with behaviour and social integration. This is the Social Stories™ method, where a story might introduce a behaviour, for example washing hands, then explaining why, how, and when a child should wash their hands. Reading the story to the children purportedly results in them subsequently enacting 'acceptable' hand-washing behaviour.

Andrews claims that through these stories, children can learn to make trait-based predictions of behaviour despite not being able to attribute mental states:

> Children who are unable to attribute mental states can come to learn how to predict behavior by attributing personality traits (…) a person with autism might be taught to associate a smile with the term 'happy', and the term 'happy' with a number of specific behaviors (e.g. hugging, laughing, etc.). The child with autism can be taught to describe a smiling person as happy, and, given that trait, predict that the person will engage in happy behaviors (Andrews 2008, p. 19).

If this is indeed the case, then this would present evidence of the possession of a theory of a trait concept that is not constituted (in part) by self-knowledge of one's own mental states; hence, it would be evidence against my dependence claim.

4.2.1 Responding to the Social Stories™ Objection

There are reasons to be sceptical of both Andrews' claim and the veracity of the data. Sansosti, Powell-Smith and Kincaid (2004); Ali and Frederickson (2006); Kokina and Kern (2010); Test et al. (2011); Nichols et al. (2015); and Wright et al. (2016) all offer reviews/meta-analyses of studies on Social Stories™. None of the cited studies in these reviews mention teaching trait-terms.

Additionally, whilst the evidence for the Social Stories™ method is positive, it is far from established. Many of the studies mentioned in the meta-reviews are noted to have professional issues, for example: citing anecdotal results, not controlling for biases, mixing other intervention methods into the stories method, and demonstrating a lack of anonymity (Sansosti et al. 2004, p. 201; Ali and Frederickson 2006, p. 371). Furthermore, Ali and Frederickson noted that "there is limited systematic evidence on the effectiveness of the social story approach" (2006, p. 356), and Sansosti et al. noted that "it may be premature, based on the current literature, to suggest that Social Stories are an evidence-based approach" (2004, p. 200).

Whilst the positive results for the method are encouraging (many studies had weaknesses, but they generally reported degrees of positive findings), it does not aid confidence in the specific claim that this method can evidence trait-based behaviour prediction. Likewise, as Gray and Garand (1993) noted, the stories require regular reading; otherwise, the children revert to their previous behaviours.[98] This suggests that there are limits to what these children are learning, and particularly that there are limits to whether they possess a theory of traits.

Let us suppose that the stories do reliably elicit the behaviour that they are trying to inculcate. Granting this, however, we might wonder the following: a) are the children engaging in these stories really those that cannot attribute mental states? If so, b) are these children really using the same concept as those with abilities to attribute mental states? Of a) the degree to which participants in Social Stories™ studies can attribute mental states has not been definitively established. Certainly, high-functioning autistic people can attribute mental states (Duverger et al. 2007; Senju 2012—though this should be obvious), so if the children involved were high-functioning, then being able to use trait-reasoning is no problem; clearly, the dependence claim is upheld.[99] However, perhaps this is meant to be a developmental claim, as the capacity for mental-state attribution generally develops more slowly for autistic children (Baron-Cohen, Leslie and Frith 1985; Kazak, Collis and Lewis 1997). Supposing that the children involved with this intervention method are young enough to have no ability to attribute mental states (I remind the reader here that there are, in fact, no data on this), we can still ask: b) Are these children using the same concept as those with the ability to attribute mental states, or even a concept at all?

For example, Andrews notes that the relevant trait-term in the stories is associated with a set of specific behaviours. Predicting one of these behaviours when children assign a guided trait-label to some other specified behaviour does not appear, to me, to demonstrate a grasp of a concept so much as a grasp of associations between

[98] The author, Gray, is the same Gray who created the Social Stories™ method.

[99] Indeed, 'high-functioning' might not even be required – there is a *world of difference* between finding mental-state attribution difficult, unreliable, and non-automatic (as many autistic people find), and literally not having the capacity to attribute mental states.

demonstrated behaviours. Though Andrews takes this to be the case—"it is a theory of behavior, not a theory of mental states" (2008, p. 19)—the breadth of this theory of behaviour is important for whether these children have a concept that could successfully refer to traits. If these children demonstrate flexible trait-reasoning in wide-ranging scenarios, then this might be a reasonable conclusion. However, as I have already noted (p. 9), without repeated story exposure, the children often revert to their previous behaviours; it is unclear whether they have successfully acquired the relevant concepts, rather than simply using terms as directed. This point can be elaborated upon using the following analogy.

Take someone shifting between slides which are various different but similar shades of blue. After repeating this a few times, she might be able to 'recognise' or distinguish a particular shade of blue from the set, after it has been shuffled. However, without repeated slide exposure, it is unlikely that she could identify that shade again (after shuffling), and would struggle to identify real-world instances of that shade. The thought, here, is that she does not have the *concept* of that particular shade of blue and thus cannot identify it out of the context in which she learned to discriminate it from other shades. In the same way as the person in this example has a concept of 'blue', but not of the shade, these children may have the concept of associations between intentional behaviours but not that of trait-behaviour—their success in making trait predictions may be down to following the doctrines of the story, rather than learning to use trait concepts.[100] Therefore, even if this particular intervention method demonstrates functionally identical results to trait-reasoning (of which there is no empirical evidence), evidence of attribution in wider contexts is required before we can accept that this method does evidence learned trait-reasoning, rather than task-specific associative reasoning.[101]

As such, Andrews' objection is based on a teaching method that has been shown not to be particularly efficacious or well-documented, but even if it were so, it is unlikely that the children involved *literally* could not attribute any mental states at all. Even granting

[100] One might respond that she has a concept of 'shades of blue' but not '*that* shade of blue'. Analogously, the children might have concepts of traits but not *that* concept of traits. However, the ability to discriminate shades is a prerequisite of the task. Assuming that the children have a concept of traits would beg the question and create a disanalogy. An analogous prerequisite might be one of knowledge that intentional behaviour has motivations, but not *trait* motivations, in which case my point still stands.

[101] Additionally, in Andrews' given example of children predicting happy behaviours, Westra pointed out that "the autism intervention Andrews describes does not seem to be about trait-reasoning at all, but rather reasoning about emotions" (Westra 2018, p. 1226).

the supposition that they could not, it is unlikely that they possessed trait concepts rather than merely having facility with task-specific associative reasoning.

In this section, I claimed that a possession condition of a trait concept includes knowledge that traits are the sorts of properties that we might have. I argued that such knowledge is derived from our ability to attribute mental states to ourselves, and showed that regardless of whether trait concepts come to be possessed by observation or by theory, the relevant self-knowledge is still required. Hence, trait concepts depend on capacities for attributing mental states, and so the dependence claim is true. I then defended this thesis from the claim that there are empirical data that demonstrate trait attribution without the capacity for mental-state attribution—I responded that the objection fails on the grounds that the empirical evidence is suspect, and even if it were not, it would not show the attribution of trait concepts. I take it that I have covered the case for trait attribution's dependence relation to mental-state attribution, but character reading is about character-based *reasoning* as well as *attribution*; as such, I move to the final section, in which I defend the dependence claim as it pertains to trait-based reasoning.

5. Conception vs. Practice: Dependence in Action

This section presents a case study to demonstrate how the dependence relation obtains in character reasoning. I will detail Kristin Andrews' pluralist account of folk-psychological explanation, and then demonstrate that it does not account for explanations of behaviour based on character. I will then argue that the truth of the dependence claim plugs this explanatory gap; hence, the recognition of the dependence claim shows us how trait explanations can be explanatory.

5.1 Andrews on Folk-Psychological Explanation

Andrews (2012) gives an account of folk-psychological explanation which, as opposed to traditional scientific explanation, does not aim for truth but explanatory coherence. This is because, for everyday social cognition, we are not typically concerned with truth so much as what makes *sense* to us. Such an account, with its focus on everyday interactions and the varieties of sense-making that we find coherent in explanation, is a pluralist account. For example, certain folk-psychological practices, such as mindshaping or stereotyping, might not require a goal of truth. Another example showing that truth is not so important, as Andrews notes, is the reality of the cognitive dissonance bias, where we are uncomfortable with information that does not fit what we understand; the existence of the bias may be used as evidence that coherence is sometimes more important than truth (Andrews 2008, p. 21). Indeed, that we aim for coherence rather than strict truth is empirically demonstrable. Andrews' account is underpinned by the work of psychologist Bertram Malle. Across many studies, Malle (2004) noted and categorized the kinds of explanations we use in everyday social interactions, finding that people's explanations could be categorised thus:

- Reason explanations: These are explanations of others' behaviour that are couched in the reasons that others had when acting in such ways.
- Causal history (of reasons) explanations: These are explanations of behaviours that are not the reasons for actions themselves, but are causal factors in the background of the reason-making.
- Enabling factors: These are the considerations that enabled the behaviour to occur.[102]

These three kinds of explanation may track truths to a certain extent, but the point is that they are different kinds of explanations of behaviour, yet each provides a kind of explanatory coherence that may be acceptable to the hearer. Using mindreading to explain behaviour, for example, would be a *reason* explanation, as beliefs/desires (and many other mental states) are reasons for action. It is notable that, in Malle's studies, only 61% of the explanations given were reason explanations, thereby demonstrating a plurality of the kinds of explanation we typically give (Andrews 2008, p. 20). Crucially, trait attributions are *causal history explanations* because they point to "factors that lay in the background of the agent's reasons", but do not cite reasons themselves (Malle, 2004, p. 91).

Despite endorsing Malle's account, it is not clear whether Andrews explicitly endorses trait attributions *qua* causal history explanations. I find this likely, however: she notes in an earlier work that trait attributions "offer explanations of behavior, even though it [the attribution] does not constitute a description of the reasons or the goal motivating the actor's behavior" (Andrews 2008, p. 22). Still, it is unclear how Andrews thinks

[102] Usually, enabling factor explanations are given when the 'why?' of the action is already known, but one wants an explanation of *how* a particular action came about, such as succeeding because of one's skill (Malle 2004, p. 10). This differs from the causal history of reasons explanations in that the causal history explanations cite some explanation that led to the target's reasons for doing something: "these explanations literally describe the causal history of reasons, which could lie in childhood, in cultural training, in personality traits, or in a situational cue that triggered a particular desire" (Malle 1999, p. 32).

traits, as causal history explanations—and not a shorthand for beliefs/desires—provide explanatory coherence. Here, I attempt to extricate an account of this from her work:

> (…) in some cases explanations in terms of trait attribution will be satisfactory; if you know about something common to all juvenile delinquents, that trait attribution might help you to understand why the child behaved as he did. Knowing that Jack is a juvenile delinquent helps us decide how to respond to his behavior, for example. Trait attribution can serve as an explanation when it yields additional information (…) by knowing that Jack is a juvenile delinquent, you can infer that his reason for stealing the candy probably wasn't to test the store's security system (Andrews 2008, p. 22).
>
> (…) other times it is clear that we do not want any more in the way of an explanation—for example, when one explains a hated politician's speech by saying 'He's either an idiot, or evil, or both' (ibid).

In the first instance (ignoring the sense in which it is highly unclear how 'juvenile delinquent' is a character trait), the trait explanation is explanatory because the attribution points to one's knowledge constituted by one's theory of juvenile delinquents—this, in turn, provides information for reasoning about the reasons the target may have had. In the latter case, the idiot/evil attribution is explanatory in that although it does not constitute a reason-giving explanation, the attribution points to further explanations of how and why idiots and evil people act the way that they do; thus, the explanation is coherent. As such, it seems that Andrews implicitly treats trait attributions, at their minimum, as being explanatory in so far as they function to *point to explainers* (such as the reasons for why juvenile delinquents and idiots act as they do).

In the pluralist spirit, I want to endorse some form of explanatory coherence account of folk psychology, especially given that the empirical data suggest that people commonly *do* give different kinds of explanations that are coherent in different ways. As such, I would prefer to use Andrews' account if possible. However, what is missing from her account is an explanation of how such trait attributions are able to function as pointers to explainers, especially given that they are portrayed here as being potentially bereft of information in themselves. This is notable when she states that "explanations that make sense are those in which the details are either filled in for us, *or those where we can fill in the blanks"* (Andrews 2008, p. 22, emphasis added). I take it that filling in the blanks would not be a problem if trait attributions were shorthand for mental-state attributions—the trait attribution pointing to the explanatory content would be handled by the shorthand relation. As she rightly argued, however, trait attributions are not shorthand for mental-state attributions.

In this subsection, I detailed Andrews' pluralist account of folk-psychological explanation and noted that it is missing a crucial explanandum, that is, how trait attributions (as causal history of reason explanations) are capable of providing the relevant inferential link to reason explanations in order to satisfy explanatory coherence. I noted that I am sympathetic to the pluralist account of folk-psychological explanation, and so it is in my interest to plug this explanatory gap, preferably in a way that demonstrates why the dependence relation holds for trait-based reasoning as well as attribution. I turn to this now.

5.2 The Dependence Relation Plugs the Explanatory Gap in the Explanatory Coherence Account

We can use what we know about the dependence claim to demonstrate how trait-based explanations of behaviour can be explanatory regarding the coherence account of folk-psychological explanation. If the dependence claim is true, then as trait explanations are constructed from a knowledge base (of traits) that are constituted in part by mental content concepts, such attributions will be conceptually richer than they were supposed to be by Andrews. This allows for such explanations to be premised on the understanding of the possibility that there *can be* reason explanations associated with such causal history of reasons explanations, as such causal history of reasons explanations are partly constituted by the knowledge of certain kinds of states that give rise to reason explanations; having facility with these extra concepts provides an explanation for why it is that trait explanations can point to reason explanations to begin with, why they are able to fill in the blanks, and why it is those reasons (and not others). Hence, the dependence relation plugs the explanatory gap in this pluralist account of folk-psychological explanation.

Indeed, it seems that Malle endorses some version of the dependence claim in his own account. Malle states that "the concept of intentionality and its critical mental state components of belief, desire, and intention lay the foundation for people's folk explanations of behavior [that include trait explanations]" (Malle 2011, p. 309). I depicted Andrews as conceiving of trait explanations as providing explanatory coherence by yielding further folk-psychological reasoning. I questioned how it was that trait explanations could point to further reasoning, if it was acceptable that they were potentially bereft of information themselves. I posited that due to the truth of the dependence claim, trait explanations should contain relevant information that enables reason explanations. When we look at Malle's account of folk-psychological explanation, we can see that this is the case: "CHR [causal history of reasons] explanations do not deny that the agent had reasons to act; they just do not directly refer to those reasons" (Malle 2011, p. 318). A trait explanation, according to Malle, need not explicitly reference a reason, such as a belief or desire, but the attribution is based on the understanding that there can be reason explanations generally, and that there *are* reason explanations which are associated with such causal history of reasons explanations.

The upshot, then, is that if we endorse a pluralist account of folk-psychological explanation, which relies on explanatory coherence rather than truth as its main aim, we find that the truth of the dependence claim helps us to explain how trait-based explanations are indeed explanatory; they are explanatory in a way that Andrews' account seems to deny the resources for when denying both the shorthand claim and the dependence claim.[103] As such, in this section I have argued that trait-based reasoning, as well as trait attribution, involves the dependence relation; hence, the dependence relation does obtain between mindreading and character reading overall, and not just for attribution.

5.3 Maintaining the Primacy of Mindreading

We have seen, then, that although a shorthand relation does not obtain between mindreading and character reading, a dependence relation does obtain regarding both the trait attributions that we make, and in at least this case study of trait-based reasoning that focusses on explanations by using traits. The dependence claim protects the primacy of mindreading (at least in relation to character reading), as it holds that there is an important constitutive relationship between the concepts applied in mindreading and those in character reading. However, it does so whilst being coherent and sympathetic to the pluralist project, in the sense that a) it accepts that character reading is a distinct folk-psychological skill from mindreading and b) that one can accept a pluralist account of folk-psychological explanation.

Conclusion

[103] I take it for granted that Andrews would want to deny the dependence claim, on the grounds that it preserves the primacy of mindreading, but of course—as is the point of this chapter—she could, in principle, endorse it and maintain a pluralist stance.

This chapter has detailed the relationship between mindreading and character reading as one of constitutively conceptual dependence, which challenges the folk-psychological pluralist critique of the primacy of mindreading. However, I have also demonstrated that mindreading and character reading are importantly distinct (by denying the shorthand claim). The key takeaway is that trait attribution, and trait-based explanations of behaviour, conceptually depend on the content attributed in mindreading. In the final chapter, I will argue for a particular ontology of the character reading cognitive mechanism; its relevance to the material in this chapter is that it will demonstrate a capacity for character reading which constitutively depends on the *capacity* for mindreading, beyond the *content* dependence established here.

From Mindreading to Character Reading: The Cognitive Gadget Account

1. Introduction

In the previous chapter, I promised an argument for the capacity of character reading constitutively depending on mindreading. In the introductory chapter, I also promised to make explicit the impact of culture on socially cognitive skills like character reading. I fulfil those promises in this final chapter by way of my argument that, inferring to the best explanation of the available information, the trait-attributive component of character reading is what Cecelia Heyes calls a 'cognitive gadget': a cognitive mechanism shaped by cultural, as opposed to genetic, evolution.

Section 2 explicates the cognitive gadget thesis. Section 3 sets the scene for trait attribution as a cognitive gadget and constrains Heyes' theory. Section 4 presents the positive case for trait attribution as a gadget. Section 5 details and responds to objections raised by a potential rival account.

2. Cognitive Gadgets

The cognitive gadgets theory (Heyes, 2018) is an empirically informed account of what makes humans such "peculiar" animals. Heyes argues that humans are similar to other animals but have developed crucial pro-social biases alongside enhanced general-learning abilities; these work together to create human cognitive mechanisms. These 'gadgets' are built in response to natural selection acting on culture: 'cultural evolution'. In Heyes' vision, genetic evolution is thus not "the architect of the human mind" (Heyes, 2018, p. 12). In this section, I give more detail on how the account purports to work and what it purports to say.

2.1 Cultural Evolution

There can be change over time in those characteristics of a population that make it distinct from others. The study of cultural evolution is the study of that change, where the relevant characteristics are 'cultural'. Here, the cultural is "the aggregate consequence of many episodes of social learning" (2018, p. 36). There are multiple ways to approach cultural evolution; Heyes takes a 'selectionist' approach. This approach makes the strongest possible claim, in that it treats cultural evolution as more than an account of how culture interacts with genetic adaptation, which is known as gene–

culture co-evolution (Feldman and Laland, 1996); it conceptualises cultural evolution as a force in its own right. As Darwinian natural selection acts upon genes, selectionists treat natural selection as also acting on culture.[104] Cultural inheritance is also subject to selective variation: Heyes notes that "the mechanisms for introducing variation — the cultural equivalent of mutations — are generators of "error" or of "innovation" in social learning" (Heyes, 2018, p. 33). In this chapter, I share an endorsement of a selectionist account of cultural evolution, but all that is really required for my account is that obtaining culturally inherited information can shape our cognitive mechanisms; a concrete example of this will be given in 2.2.

2.1.1 Cultural Evolutionary Psychology

Cultural evolutionary theory focuses on what Heyes terms the '*grist*' of the mind; that is, on the behaviours and "conceptual structures" (societal norms, et cetera) that populations express in the cultural domain (pp. 36-37). Heyes' work examines how cultural evolution might act on cognitive mechanisms, on the '*mills*' of cognition that allow the production of behaviours/conceptual structures. Previously, it had been assumed that while behaviours are subject to cultural evolution, their mechanisms for action must have genetic bases (p. 14). Heyes argues that cognitive mechanisms can be products of cultural evolution too; as such, many cognitive mechanisms are 'inherited' and developed during the lifetime of the human through the aggregate consequences of social learning. This is not to claim that whole cognitive mechanisms are copied from one brain to another through culture, but that, for example, "certain kinds of social interaction, sometimes with many agents over a protracted period of time, gradually shape a child's cognitive mechanisms so that they resemble those of the people around them" (p. 44). To resemble here is to share parts that have been adapted from existing genetically coded cognitive structures, such that pro-social biases and powerful learning mechanisms use existing cognitive architecture to tweak and change how and for what purpose these mechanisms work.

Whilst cultural evolutionary psychology represents a new angle for the psychological study of the mind, noting how it shapes our brains as well as our behaviours, the cognitive gadgets theory goes further. I outline two extra claims that Heyes' book on gadgets makes below, though one should note that—in my constraining of the gadgets

[104] Strictly speaking, selection acts on phenotype—I say genes here to contrast this account against cultural accounts that do not require multi-generational genetic involvement.

account in Section 3—I do not commit to an endorsement of the truth of these following two claims.

2.1.2 The Ubiquity of Gadgets and its Distinctiveness for Humans

The first additional claim is that of the ubiquity of gadgets. This is the claim that not only are cognitive gadgets real, but they are also ubiquitous in the mind. That is, the main developmental structure of our cognitive mechanisms is such that they are culturally acquired and built upon a basic genetic starter-kit of pro-social biases and enhanced learning apparatus.[105] This is part of what makes Heyes' account a radically constructivist account; it construes the mind as essentially learning and building all of this cognitive machinery itself without the aid of genetic code for building that machinery. Her argument for this is given across her book in consideration of four target capacities, which she argues are gadgets: social learning, imitation, mindreading, and language. The extent to which these four capacities are captured accurately by constructivist accounts is contested, but if these four *are* gadgets, then it becomes difficult to deny that such an organisational structure of the development of such capacities is likely to be different for other cognitive mechanisms.

The second additional claim is that the cognitive gadgets theory provides an explanation for why humans are such 'peculiar' animals. That is, the construction of a suite of cognitive gadgets is responsible for what makes humans so noticeably different from other animals. That said, Heyes notes that "I am not saying that distinctively human cognitive mechanisms must have been shaped by cultural evolution" (p. 77). The point is that there is a traditional question of 'nature vs. nurture' as it pertains to psychology. Of course, some human capacities, which are not otherwise genetically encoded (nature), will be products of nurture (of learning), but Heyes' project aims to show that nature, nurture and *culture* exist together to form *human* capacities: "Culture is 'nature-like' as a product of a selection-based inheritance system, but 'nurture-like' in being acquired in the course of development through interaction with the (social) environment" (p. 45).

What we have in the cognitive gadgets theory is an examination of how cultural evolution may affect the construction of cognitive mechanisms, not just behaviour. Heyes' book (and to some extent her prior papers on the subject, for example Heyes,

[105] I say more about this learning apparatus in 3.2.

2012a, 2012b) argues that gadgets are ubiquitous in the brain, and that their varied construction is what makes humans so distinct from animals, or rather, peculiar *as* animals. What is still required, though, is a clear example.

2.2 Literacy as a Gadget

A suitable example would be one in which the behaviour is distinctively human and is culturally acquired, with evidence that the cognitive mechanism that gives us the capacity for the behaviour is acquired without genetic inheritance. A compelling case is our capacity for literacy. Literacy is obviously distinctively human and is culturally acquired. It takes a great deal of input from pedagogues, informed by culture (the alphabet, grammar, pedagogic methodology, et cetera), for the subject to gain the ability. The behaviour, the reading of a particular language, is culturally inherited, but there is also evidence that the cognitive mechanisms responsible for it are too.[106]

The focal area of the occipitotemporal cortex activates in brain scans much more strongly in literates than in illiterates; this area of the brain is known as the 'visual word form area' (Heyes 2018, p. 20). This area of the brain tackles object recognition generally, but when the subject becomes literate it rewires into the visual word form area to process the shapes of written words—it is responsible for processing representations of the lines that make up alphabetic symbols. Crucially, literacy has arisen in human populations too historically recently for there to be genetically encoded systems for it; hence, the visual word form area is not a mechanism produced by genetic evolution (Dehaene and Cohen 2011). The rewiring of the object recognition centre into one that specialises in recognising word forms and facilitating the recognition of words thus warrants an explanation—one that a gadget account provides.

That said, there are two critiques that we might make of literacy as a gadget. The first is an empirical challenge that the visual word form area is not specialised for word form recognition; the second is a theoretical challenge to the supposition that if some

[106] It has been suggested that perhaps 'intelligent design' (the structured improvements of informational understanding and its inculcation, for example literacy programs) plays a role in gadget construction, as well as cultural evolution (Morin, 2019). This is perhaps so, given that literacy cannot be acquired without specialised training, and much culturally inherited information will have been structured for optimal uptake. However, Heyes notes that no one has yet produced efficacious programs for the development of mechanisms, rather than mere behaviour (Heyes, 2019, p. 3). Furthermore, Morin's claim is that intelligent design might serve *alongside* cultural selection as a developmental force; this is not incompatible with any of my claims.

capacity is not the result of genetic adaptation, then it must be a gadget. I will respond briefly to these criticisms now.

2.2.1 Critique of the Visual Word Form Area

The visual word form area has been challenged in its designation of specifically computing the form of words (Price and Devlin 2003). This is because the area activates when it is processing for other tasks, such as naming colours, and representations of a word form recruit the use of other cortical areas besides the visual word form area. However, it does not follow from this that the visual word form area *has not* adapted to compute word forms during literacy acquisition. The visual word form area does process other tasks, as might be expected for engaging in the tasks it was used for prior to word form recognition, but once literacy has been achieved the area does light up under functional magnetic resonance imaging scans more strongly for word form processing than anything else (Dehaene and Cohen 2011, p. 257).

Secondly, the debate somewhat rests on what one means by 'specialising for a task'. On an overly strong view that takes cognitive specialisations to be confined to one cortical area, we might conclude that the visual word form area is thus not specialised for word form cognition, as other brain areas are involved. However, even strictly modular accounts of cognition need not commit to a particular neural locale for the entirety of a mechanism's domain-specific functioning. It seems more reasonable to conceptualise specialisation functionally, along the lines that if most of the work of word form processing is performed in the areas defined by the visual word form area, then it is still of theoretical interest in the causal story of literacy acquisition.[107]

Indeed, one of the main pieces of evidence for the visual word form area being responsible for word form processing is that a lesion in the visual word form area appears to result in alexia: the loss of the ability to read due to the inability to parse the letters on the page. That said, Price and Devlin rejected this on the basis that alexia is a general visual problem; it can be invoked in a multitude of ways. Furthermore, they argued that claims of complete alexia resulting from lesions in the word form area were unfounded (pp. 474-475).

[107] See Chen et al. (2019) for a discussion of the brain circuitry underlying the visual word form area and its relation between word form and attentional processing. This goes some way to filling the gaps implied by Price and Devlin's criticisms of what and where the visual word form area computes.

Cohen and Dehaene (2004) responded that of course no lesion was likely to be so perfectly disruptive that it resulted in pure alexia, but the word form area theory predicts the results of partial lesions quite well. If the word form area processes the shapes of words, then imperfect lesions on it should impact the cognition of word shapes in imperfect ways. They note:

> For instance, the patient studied by Miozzo and Caramazza (1998) was severely impaired at naming single letters, and she was unable to decide whether an upper- and lower-case letter had the same name or not. However, she could accurately discriminate real letters from visually equivalent pseudo-letters, as well as normally oriented letters from mirror-reversed letters (Cohen and Dehaene 2004, p. 472).

In this case, a lesion on the word form area resulted in specific dysfunction in the processing of word forms, whilst other aspects of processing word forms that required some kind of contextual knowledge about word shapes (that such a shape does correspond to a real letter, for example) were unaffected. As such, despite the criticisms of the categorisation of the visual word form area as being a specific area of the brain specialised for word form recognition, the data are consistent with the claim that the visual word form area is a powerhouse of the processing of objects and lines that specifically rewires to be very good at detecting the lines and shapes of letters and words for reading; this occurs as one becomes literate, and so cannot be the result of genetic adaptation. The claim is that it is instead a cognitive gadget. But are there other options?

2.2.2 What Influences the Developmental Structure of Cognitive Mechanisms?

Whether there are other options between genetic encoded structures and cognitive gadgets depends on what can affect the structure of cognitive mechanisms. I propose that there are two relevant options to consider. The first is epigenetic expression. The second supposes that there are mechanisms for learning certain capacities that are genetically encoded, such that the mechanism bootstraps the acquisition of the capacity. This latter case is dealt with in Section 5, in which I pose and respond to a rival account to the gadget thesis. I see this as the main explanatory competitor, and so it receives more detailed treatment later when my account of trait attribution as a gadget is clear.[108] Of the former case, there is indeed evidence that epigenetic processes can change the

[108] However, such an account cannot lay claim to literacy, given that we know that literacy has not been genetically encoded at all.

cognitive structure, but this is consistent with how gadgets might be built. I briefly discuss epigenetics in order to demonstrate this.

Epigenetics deals with changes in how genes are read or expressed, which do not alter the DNA itself.[109] This functions usually to regulate gene expression, but also has a developmental function in regulating certain states of neuronal activity (Lux, 2013, p. 75). It is, in simple terms, the biochemical mechanism by which our bodies adapt to our environments, allowing for certain genes to be switched on or off in order to produce or stop producing certain proteins. Generally, protein production is regulated by such processes as DNA methylation, where methyl groups are added to the DNA sequence, like adding notations to sheet music (Bommarito and Fry, 2019). These changes are usually thought to be wiped clean between generations, as DNA is demethylated when it is transferred from parent to offspring (Monk, 1995).[110] As such, epigenetics provides a way of effecting brain biochemical changes in response to environmental factors that are generally not passed on to offspring. Could epigenetics, then, be an alternative method of acquiring literacy that is not genetic adaptation?

The answer is yes, but this is coherent with a gadget thesis, not an alternative explanation. There seems to be a developing understanding that DNA methylation has a relationship with certain cognitive developmental disorders (Lalande and Calciano 2007; Coppedè 2014; Mendiola and LaSalle 2021), with scholars also noting a relationship between epigenetics, learning, and the construction of memory (Day and Sweatt 2011; Marshall and Bredy 2016). As such, epigenetic expression appears to be an important biochemical mechanism for effecting structural changes to cognitive mechanisms. This does not mean that such changes can be achieved without input from the environment, such that literacy could be acquired *solely* through epigenetic means.

[109] Deans and Maggert (2015) note that the term 'epigenetics' is somewhat confused in the neuroscientific and genetic literature; there are two research programs based on two definitions: "Waddington's definition is largely used to describe the expression of environmentally mediated phenotypes (…) Those in the field of genetics concerned with DNA methylation, chromatin activity states, chromosomal imprinting, centromere function, etc., predominantly use Holliday's notion of epigenetics. They are interested in how expression patterns persist across different cells (mitosis) and generations (meiosis)" (p. 889). The environmental mediation of phenotypes is nonetheless achieved through the molecular processes that Holliday's adherents are interested in, so my discussion of epigenetics incorporates these by describing epigenetics as being possibly environmentally mediated gene expression, which takes form at the molecular level in processes such as DNA methylation, et cetera.

[110] Despite this, there are some studies which are beginning to report transgenerational epigenetic inheritance (Bond and Finnegan 2007; Wei, Schatten and Sun 2015). However, the strength of the transgenerational inheritance is rather weak; such changes only persist for a couple of generations, and are confined to specific examples.

Indeed, the relationship between learning and memory suggests a speculative but plausible explanation for the biochemical construction of cognitive gadgets resulting from the inheritance of cultural information. In this sense, epigenetic expression is no threat to the gadget thesis.

With such a defence, the claim that literacy is a gadget is consistent with the empirical data and supported by theory. This section has explicated the cognitive gadget thesis, given an example of a paradigmatic cognitive gadget, and defended that designation against the claims that the relevant cognitive mechanism does not do what it says it does, and against the objection that there may be other ways of enacting relevant brain changes besides a dichotomy between genetic adaptation and cultural adaptation.[111] I now move to motivating the issue at hand: the consideration that trait attribution might be a cognitive gadget.

3. Considering Trait Attribution as a Cognitive Gadget

The dependence relation I established in Chapter five gives us pause to consider the ontology for the capacity of trait attribution. A conceptual and explanatory dependence on mindreading by character reading has been established—could this dependence extend to the cognitive machinery that grants trait attribution as well? I will argue in 3.1 that the developmental data on the acquisition of the capacity are consistent with trait attribution as a gadget, though more evidence is needed. In 3.2, I will constrain the gadget thesis on the basis that trait attribution could be a gadget independently of the truth of the ubiquity-of-gadgets claim, and independently of the human distinctiveness claim.

3.1 Developmental Psychology of Trait Attribution

Andrews (2008, pp. 18–19) provides some detail on the development of trait-attributive abilities. She notes that “When a child is able to talk about and respond to others’ beliefs and desires at age 3 1/2, she is still unable to use traits to predict behavior (Kalish, 2002; Rholes et al. 1990)”. This suggests evidence of character reading which is apparent after mindreading emergence. Furthermore, Yuill (1997, p. 281) found that children start making limited predictions of behaviour based on traits at four years old, but become much more advanced by age six/seven. If these considerations are evidence that character reading develops later than mindreading, whilst depending on the

[111] Though the second half of this argument is given in Section 5.

mindreading ability in the ways detailed in the previous chapter, then this would certainly motivate a claim that trait attribution is a gadget built on mindreading architecture.[112]

Unfortunately, the picture is not so clear. The data on trait attribution developing after mindreading have been challenged by Westra (2021) on the grounds that the bar for trait recognition was set too high in those studies—the evidence suggests that the abilities first manifest at around the same time (p. 8221). Granting this for the sake of argument, we are left in a situation in which the ability to character read seems to develop at the same time as the ability to mindread—where does this leave us regarding my claim that the capacity for character reading ontologically depends on mindreading? Furthermore, how does this relate to the gadget claim?

A claim that trait-attributive systems develop as a gadget, and therefore that our socially cognitive mechanisms are shaped to resemble others, only requires that there was no *prior* trait attribution. The evidence of trait attribution and mindreading being detectable at the same time in development is not yet accurate enough to determine *precisely* when one ability is first detectable and is consistent with a claim that the ability develops as a result of pro-social biases and enhanced learning methods. Indeed, at the point where mindreading and character reading skills develop explicitly, children's understanding of traits appear to be mentalistic in nature, such as their understanding being premised at least in part on an understanding that others can have subjective and competing desires (Yuill and Pearson 1998, experiment 2); hence even if they develop together, trait attribution is reliant on mindreading structures.[113]

However, Yuill and Pearson do claim that "The results suggest that children change from viewing traits as behavioral regularities to understanding them as internal mediators, and that advances in understanding desire underlie this change" (Yuill and Pearson 1998, abstract). As such, there is a question over whether traits conceived of as behavioural regularities are to be considered as true traits, or as something else. For my

[112] There are two claims that come apart here: 1) that trait attribution is a gadget, and 2) that trait attribution is a gadget built on mindreading architecture. It is certainly possible to be the first without the second, but I entertain the second possibility for the following reason: if mindreading is a gadget, then there is no reason in principle why gadgets cannot be built on other gadgets, given their propensity for the use of existing neural hardware to construct themselves. If mindreading is not a gadget, then it has a genetic basis, and therefore gadgets may be built on the genetic base, whatever that may be. Either way, trait attribution can be a gadget regardless of the status of mindreading.

[113] This evidence is also supported by my theoretical arguments, given in Chapters two and five, about the nature of trait concepts.

purposes, I will rely on the material in Chapter two that argues for why trait concepts are psychological concepts. However, I do not think that this contradicts the story of the developmental data from Yuill and Pearson; it seems we disagree over the nature of the concepts involved in trait tracking. This is because on my understanding of what traits are, argued for in Chapter two and defended further in Chapter five, children are not tracking character traits until their understanding of 'traits' is mentalistic. As such, there is a distinction between a behavioural trait and a character trait, the latter being the focus of this thesis. Meanwhile, Yuill and Pearson conceptualise the developmental story of trait attribution as one that includes conceptual development over time, so all trait tracking is character trait tracking (of various complexities).

Despite this difference, their account is not necessarily incompatible with the claim that trait attribution does not occur before mindreading, which is a key takeaway of this subsection. This is because the sense of desire that Yuill and Pearson work with has undergone conceptual development already: it goes beyond a simplistic understanding of a desire as a preference—their subjects had to recognise desires as subjective and potentially competing (p. 583). As such, if simplistic trait attribution can occur before full-blown mindreading, then this data is consistent with a claim that simplistic mindreading could occur before full-blown mindreading (and hence one cannot decisively claim that trait attribution can occur before mindreading).

Indeed, Yuill and Pearson's work concerns explicit paradigms for trait attribution, but there is also data on implicit methodologies to consider. Westra (2021, p. 8220) notes that whilst the nature of *infants'* trait understanding as being mentalistic or not is still contested, mentalistic proponents can appeal to studies such as (Repacholi et al. 2016) to argue that

> infants are able to form traitlike generalizations over agents' goals and over their emotional states, that they are able to integrate those generalizations with their knowledge of perception to generate different predictions across contexts, and that they are able respond appropriately on the basis of these predictions.

In comparison, the non-mentalistic camp might appeal to behavioural-association accounts, but Westra noted that the infants' capabilities in the studies could be generalised to success in new contexts, so infant reasoning in such trait-studies appears to track more than simple behaviour-reading associations (ibid). Such considerations would thereby be a route for maintaining that trait attribution and mental-state attribution are evidenced developmentally together by appealing to implicit data (where the implicit mental-state attribution data consists in Onishi and Baillargeon's studies, amongst others).

In sum, the point of this discussion is to highlight that there is no developmental data that *disqualify* trait attribution from being a gadget. This is, of course, not a great consideration alone as a motivation to think that trait attribution is a gadget. It serves as part of the consideration that the evidence thus far is consistent with the gadget thesis (such as conceptual and explanatory dependence), and thus warrants deeper critical examination. I now move to constraining the gadget thesis. Whilst trait attribution may be a gadget, it might not be true that gadgets are ubiquitous in the mind, nor that gadgets are what makes humans distinctive. I show how these issues come apart below, thereby demonstrating that I can be neutral on these topics when giving positive evidence in Section 4.

3.2 The Ubiquity-of-Gadgets Thesis is Contentious

I have argued that at least one gadget exists, implicitly assuming that there might be more. This is in contrast to Heyes' stronger claim that gadgets are ubiquitous. If gadgets are indeed ubiquitous, then it is highly likely that trait attribution is a gadget. On such an account though, mindreading is a gadget. I do not want to put the cart before the horse, considering that the ubiquity claim is contested on account of its presumed developmental account of psychological capacities. Below, I briefly outline the constructivist and competing nativist account of how psychological capacities come online, in order to show that I can remain neutral on the question of whether mindreading is a gadget whilst arguing that trait attribution is indeed a gadget.

3.2.1 Heyes' View of Cognitive Architecture

On Heyes' view, 'domain-general associative learning' mechanisms build gadgets. Associative learning mechanisms are found "in every vertebrate and invertebrate group where it has been sought, and in a wide range of functional contexts, from foraging to predator avoidance, mate choice, and navigation" (Heyes 2018, p. 68). The 'associative' learning can be defined as "learning in which an excitatory or inhibitory link is formed between representations of events"; an example would be Pavlov's dogs (p. 67).[114] Associative learning is known to be genetically inherited; even infants display associative

[114] Associative conditioning is associative *learning* when conditioned information is used functionally.

conditioning (p. 70). Furthermore, there is an evolutionary history to the capacity, where for example baboons form more associations than pigeons (pp. 69-70), and regarding advanced development in humans "the process became dependent on prediction error. For an association to be formed, a pair of events still had to occur close together in time, but, in addition, one event had to be predictive of the other" (p. 68). As such, making associations is not only about linking random representations in time but also recognising causal relations.

It is Heyes' claim that associative learning is how children master concepts (p. 71), and how one is able to build cognitive gadgets. Domain-general learning is needed for gadgets because any culturally acquired cognitive capacity would take information from many different domains and require many concepts. Any learning mechanisms that humans have, which are *specialised* for certain tasks or types of tasks, will not have the right amount of plasticity to develop gadgets.[115] The claim, then, is that this uncontroversial form of learning is controversially responsible for building most of our cognitive architecture.

3.2.2 Radical Theses are Naturally Contentious

However, Heyes' argument that gadgets are ubiquitous is given across her book in consideration of her four targets, which are all types of social capacities (social learning, imitation, mindreading, and language). When it comes to the cognitive mechanisms of these capacities, there is debate over whether they could indeed be built solely from domain-general learning processes, or whether they require some genetically encoded innate information or specially designed learning mechanisms for the required domain. Indeed, there has been keen engagement with this question for each of her proposed gadgets (Heyes, 2019a). Furthermore, as noted in the introductory chapter, there has been debate between constructivists and nativists about the correct developmental account of mindreading (Laurence and Margolis 2001; Ray and Heyes 2011; Carruthers 2013; Heyes 2014). Crucially, both sides use their accounts to accommodate the existing developmental data.

As such, whilst the existence of gadgets is not contested, the gadget status of these targets is. What would it mean for trait attribution if the ubiquity thesis was true, or false? Thankfully, it would mean nothing. If the thesis is true, then trait attribution is

[115] 'Plasticity of the mind' refers to how flexible and easy it is to accommodate cognitive changes based on novel inputs.

highly likely to be a gadget. This is independent of the question of whether one takes, as I do, mindreading and trait attribution to be different capacities, or whether one takes trait attribution to be part of mindreading (as Westra 2018 does). If the ubiquity thesis is false, then because we have evidence of at least one gadget, mindreading may well form part of the genetic starter kit of pro-social biases that are necessary for gadget construction.

3.3 The Distinctively Human Thesis is Contentious

Much the same reasoning applies to the claim that the human propensity for building cognitive gadgets is what makes us such 'peculiar' animals. In this sense, the truth of the distinctively human thesis is orthogonal to whether trait attribution is a gadget. Furthermore, it is not yet clear that gadgets are characteristic of human distinctiveness. The claim may be cut two ways: the strong claim is that the ubiquity of gadgets is what makes humans so peculiar as animals, whereas the weaker claim is that it is merely the presence of gadgets that is distinctively human. On the former understanding, we can side-line the distinctively human thesis immediately (given my noted neutrality about the ubiquity thesis). On the latter sense, we need to be hesitant because developing gadgets is not likely to be an all-or-nothing affair; a spectrum of development is likely.

If a cognitive mechanism is the product of cultural evolution, then it cannot be a mechanism that non-human animals have, as non-human animals are incapable of developing culture through aggregate episodes of social learning. Whilst non-human animals *are* capable of social learning, and perhaps capable of *aggregate* social learning, they are not capable of aggregate social learning that produces consequences of the kind (in complexity and longevity) that humans can. However, Heyes does acknowledge that some cases of human cognition which she deals with may be present "in nascent form, in nonhuman animals" (2018, p. 42). Culture begins somewhere; hence, the products of cultural evolution will begin somewhere—it is unlikely that cognitive gadgets would just spring into existence at some critical mass of cultural engagement amongst a species. Heyes' (correct) concession that such processes are not all-or-nothing seems to be too much for this weaker distinctiveness claim, though, as she considers that domain-general associative mechanisms have an evolutionary history that puts the ability on a spectrum of development. The point being that rudimentary gadgets based on rudimentary cultural inheritance would likely start to develop in animals at some point; hence, we need not commit to the second sense of the human distinctiveness claim.

As such, we can constrain the gadget thesis on the above grounds. I am merely arguing that trait attribution is a gadget, and I can argue for this without committing to a nativist or constructivist view of mindreading. Furthermore, I can argue this without accepting the ubiquity-of-gadgets thesis, nor the human distinctiveness thesis. I now turn to the positive case for trait attribution as a gadget.

4. The Positive Case for Character Trait Attribution as a Gadget

My positive case consists of two arguments. The first is theoretical, in that it is an argument based on the evolutionary development of the trait-attributive capacity. The second argument is empirical, in the sense that it takes the spontaneous trait inference mechanism as a case study and argues that the data imply that it is a gadget.

4.1 Trait Attribution Outpaces Genes

All humans have personalities; in principle, it might be selectively advantageous to recognise what sorts of people you are dealing with, and to be able to predict and explain their behaviour by reference to these categorisations. But what mechanism of selection is more likely here: genetic or cultural? In order to answer this, consider the markers of personality, such as certain gestures, facial expressions, and types of action, et cetera. Recall the austere gentleman from Chapter three, and the issue that with this multiplicity of trait expression, these types of expression can be interpreted along many different axes of trait dimensions. Not only are the markers of personality going to be highly variable within a population, but they are also subject to change across cultures *and* across individuals over time in a culture. For example, if a man is grumpy all his life, this may manifest in crying as a baby, being terse in his teenage years, being aggressive in his prime, and being anti-social in old age. Similarly, within the same culture, another grumpy person might refuse to eat food as a baby, be easily angered as a teenager, scowl and rarely smile during their prime, and be dismissive of the youth in their old age. Both the norms of behaviour and expressions of trait-relevant properties change across an individual's life, even within a singular community. A mechanism needs to develop to build and amend associations between trait markers and trait concepts—as such, a cognitive mechanism selected to track these behaviours and mental states needs to adapt as quickly as the norms of a community. Indeed, Heyes cites this tracking of "labile features of the environment" as key evidence that mindreading is a gadget (Heyes 2018, p. 208).[116]

In cases of rapid changes between the associations between concepts and behaviours, the claim is that associative learning mechanisms have the power to build and amend these associations, to biologically instil the aggregate consequences of social learning, whereas genetic adaptation is not ontogenetically flexible. Even in cases of people living in the same cultural communities for generations, this will not guarantee the genetic selection of such a mechanism. As Heyes notes, "in the stable phase, 'assimilative alleles' – genes that reduce the experience-dependence of a cognitive gadget's development – may increase in frequency. But when the environment shifts, there will be selection against assimilative alleles because their bearers will be slower to adjust to the new conditions" (pp. 208-209). So, the argument is that trait attribution outpaces genes in that what it needs to compute varies too much for genetic selection to take a hold, to build a mechanism for it in subsequent generations—the changing trait expressions of a community can be better captured by cultural selection, in which the capacity is built through the inheritance of this cultural information.

4.1.1 Objection: Some Traits are Cross-culturally Stable

One might reply, however, that some traits appear to be largely stable across cultures; take for example the emotion-related traits (being a happy/sad/angry person)—could these not be selected for? In order to attribute such a trait, one needs to understand, for example, happiness as an emotion that is a manifestation of some underlying trait. Indeed, the ability to recognise emotions from faces is a cultural universal (Ekman and Cordaro 2011). Furthermore, the interpretation of specific emotions on faces, such as happiness, or rather one of the six basic emotions (fear, anger, joy, sadness, disgust, surprise), is often also touted as a cultural universal, based on several cross-cultural studies of emotional expression (Ekman and Friesen, 1971; Izard 1971; Ekman et al. 1987; Matsumoto and Ekman 1989; Ekman 2006).

However, whilst there is some cross-cultural consistency in emotion attribution, it is not clear that recognising specific emotions from specific facial displays is a cultural

[116] However, it is not obvious that mindreading tracks features of the environment that are as labile as traits – it seems less problematic to me that some genetic selection for tracking kinds of psychological states of others is appropriate, but I shall say no more; I promised neutrality on the claim that mindreading is a gadget.

universal. Russell (2016, p. 163) noted that Ekman's (1971) study confounds facial physiognomy with expression; Kayyal, Widen and Russell (2015) demonstrated that context affects emotion perception, even across emotions with opposite valence; Gendron et al. (2014) found that the remote Himba ethnic group in Namibia reported a differing pattern of recognition for emotions compared to Westerners; Jack et al. (2009; 2012) argued that the perception of emotion was culturally variable based on the cross-cultural confusion of certain basic emotions such as surprise and disgust (particularly in East Asian cultures), and Nelson and Russell (2013) conducted a meta-analysis suggesting that culture and language mediate the recognition of emotions from faces.

The evidence thus suggests that the recognition of both emotion markers and the intensity of those emotions is not a cultural universal. Clearly, there is some cross-cultural consistency, as shown by Ekman's studies, but the fact that specific interpretations of faces as evoking X emotion are not universal (and hence emotion-related traits will not be universally identifiable) suggests that such attributions will require the tracking of the labile features of the environment; this is even if features like smiling and recognising smiling are governed by genetically adapted mechanisms.

4.1.2 Objection: All Humans can Attribute Character Traits

However, it might be wondered why cross-cultural differentiation in attributing traits means anything, on the grounds that it is obvious that all humans can attribute character traits. This reflects a general scepticism of the same kind noted above regarding the recognition of facial expressions—we can all interpret facial expressions as expressing emotions, can we not? Of course, my answer to that critique was to point out that whilst we can all attribute emotions to faces, the data suggest that an apparently obvious fact about psychology in our picking out of specific emotions was not universal. In this case, my response is in a similar vein. Just because we can all learn the practice of character reading, and thus shape our cognitive mechanisms to be like those around us (as in literacy), does not mean that it is a universal practice. In fact, there is evidence that trait attribution is historically a culturally variable practice.

It may be helpful to conceptualise this in comparison to what our folk norms of psychology are. Historically, a large set of psychological studies across all domains of enquiry have been conducted on WEIRD people—that is, people of Western, Educated, Industrialized, Rich, and Democratic cultures—despite their being a small representation of the world's diverse population (Henrich, Heine and Norenzayan

2010a, 2010b; Henrich 2020). This may mean that many pre-theoretic folk norms of *human* psychology that have sedimented into social consciousness are based on a set of data that is not indicative of all humanity.

In order to demonstrate that trait attribution has a historically variable level of importance as a folk-psychological practice, I discuss two instances: one from a WEIRD cultural set and one from a non-WEIRD culture. The first instance is in personality's historical influence in popular culture. Baumeister (1987) discussed the development of issues surrounding selfhood from a historical literature perspective, on the assumption that—much like today—literature and historical texts carry themes of the cultural zeitgeist. His work focussed on literature from France, America, and England, so variation here occurs amongst cultures we would call WEIRD. He noted that selfhood was not even a popular theme until the 16th Century (p. 165), and that:

> During the 19th century, personality (rather than social rank and roles) came to be increasingly regarded as a, even the, central aspect of the self. One source of evidence for this development is trends in biographical writing, which shifted to emphasize personal material (Altick, 1965) (p.166).

Whilst character has always been part of our conception of ourselves, its cultural *relevance* as a socially cognitive dimension of understanding others has been historically varied, even within WEIRD cultures.

The other instance, focussing on a non-WEIRD example, is one in which personality and associated character traits were just not a relevant socially cognitive factor. For example, Miller (1984) noted in his work with Indian Hindus situated in Mysore that they referred to relationships between people, rather than internal traits. In particular, he noted that "most of the cross-cultural differences observed in the use of general dispositions resulted from references to personality characteristics"—something which was more common for Americans than the Hindus in the study (p. 967). The point being that even though all of these cultures had the ability to attribute traits due to the robust socially cognitive capacities that most humans possess, the cultural impact of varying conceptions about personality affected socially cognitive practices across cultures. Indeed, the effect of enculturation can be noted especially in this study, as American and Indian children gave very similar responses to each other, but the adults differed quite starkly cross-culturally in the trait dispositional explanations that they gave (ibid).

One might note, however, that all that has been established is that trait-attributive practices have developed over time and that they vary across cultures. This is certainly incompatible with any mechanisms for them being genetically encoded, but maybe we can just learn these practices, where trait attribution is handled by some general cognitive mechanisms that we already have. To that end, I now move on to a discussion of a particular cognitive mechanism of trait attribution, arguing that this counts as an example of a cognitive gadget.

4.2 The Spontaneous Trait Inference Mechanism as a Cognitive Gadget

Previous chapters have introduced and discussed the importance of spontaneous trait inferences. Here, I think that they may be used as part of the positive evidential case for character trait attribution being a gadget. As noted, character traits are often attributed to people at the behavioural encoding stage of social interactions. What is crucial about spontaneous trait inferences is that there are no overt behaviours associated with them (as there might be with other social practices). As such, there is no confusion as to whether the behaviour elicited is a result of cultural evolution but has a genetically encoded mechanism, or the mechanism itself is the product of cultural evolution—the data only pertain to how the mechanism functions, as the trait attributions are only detectable after the fact (using memory tasks based on the encoding specificity principle, as noted in Chapter one).

It is pertinent to the thesis that trait attribution is a gadget that our knowledge of people can affect spontaneous trait inferences: their attributions are affected by psychological distance (strangers vs. friends) (Rim, Uleman and Trope 2009), and by the subject's goals (Uleman, Adil Saribay and Gonzalez 2008). If such a cognitive mechanism were the result of genetic encoding, we would expect the automatic system to function *independently of experience*—much like how a knowledge of illusions does not affect how illusions actually appear; these data show that this is not the case. Additionally, Hassin et al. (2004) showed that American children use more spontaneous trait inferences compared to non-Westerners from collectivist cultures.

For ease of example, recall the classic debate between nature and nurture, and how Heyes' work purports to support the integration of culture with these two. The fact that our *unconscious* trait attributions vary according to our cultural situation is discordant with a) a 'nurture-based' claim that trait attribution is merely a learned reasoning process with no dedicated cognitive machinery involved, and b) a 'nature-based' claim that innate capacities generally function independently of experience (because such mechanisms would be there regardless of the experiences of the subject).

As such, the fact that the trait attributions made in spontaneous trait inferences are culturally sensitive—despite being automatic—is significant, but there is more that we can say. Cognitive mechanisms are physical, so we can look at neural data to assess their status as a gadget, much as we did with the visual word form area. Data on the neural correlates of spontaneous trait inferences are sparse, as noted in Chapter three, but there are some relevant neural data on spontaneous trait inferences that pertain to when they occur.

4.2.1 Spontaneous Trait Inferences and the N400

The neuroscientific data supporting the cross-cultural variance of spontaneous trait inferences pertain to the detection of the so-called 'N400'. The N400 is a pattern of neural activation response measured by electroencephalography (Kutas and Hillyard 1980). It is, essentially, a particular kind of electrical signal of the brain detected by electrodes on the scalp. The amplitude classically peaks at around 400 milliseconds after exposure to particular kinds of stimuli regarding semantic meaning (hence N*400*), so it is generally held to be involved in the processing of meaningful (read: contentful) information; its discovery has helped us to investigate the neuro-functional relationship between perception, memory, attention, and language in comprehension (Kutas and Federmeier 2011).

Here I will outline the prediction regarding the N400 as it pertains to trait attribution, then briefly run through the relevant study in order to show that, as per the results of this study, it appears that there is cross-cultural variance in making spontaneous trait inferences. Given that the N400 is an event-related neural response to the processing of meaning, I claim that we should expect N400 activation if a spontaneous trait attribution is made. Crucially, though, as will be noted in the following discussion of the relevant study, we should expect *strong* N400 activation if a spontaneous trait inference is made but then the subject is presented with stimuli which are *incongruent* to the trait inference (such as words or faces representing incongruent traits).[117]

[117] This is due to the fact that repeated words, and similar words with similar meanings, illicit a weaker sequential N400 amplitude, whilst unexpected words elicit stronger frequencies (Kutas and Hillyard 1984). One should also note that the N400 is not merely a response to *unexpected* stimuli, as this was controlled for in the original study (Kutas and Hillyard 1980, p. 204). Relatedly, Van Duynslaeger et al.

Na and Kitayama (2011) tested this prediction by conducting a study on two cultural groups recruited from the University of Michigan— 'European-Americans' and 'Asian-Americans'.[118] In study one, in a memorisation phase, they paired faces to behaviours that jointly implied certain traits. The participants were not instructed to make a trait attribution, so if one were made it would be spontaneous. Then, they presented the targets with a task that presented a face from the memorisation trial, with either the trait implied by the face, an incongruent trait, or a pseudoword. Participants pressed one of two keys to say whether the word was an English word or not.

The initial prediction was that European-Americans would make spontaneous trait inferences from the face/behaviour pairs and thus be more *accurate* in the relevant trait conditions than in the incongruent trait conditions, whilst Asian-American participants should show no difference in accuracy as their responses would not be primed by previously made spontaneous trait inferences (p. 1026).

When participants were presented with faces that were incongruent to the traits implied by previous face/behaviour pairings, they indeed found this effect to occur (p. 1027). Accuracy, however, does not tell us a great deal, but they also calculated *response times* to pressing the button. They found that European-Americans were much slower to press the button in the incongruent trait condition, whilst Asian-Americans were basically identical in response timings on that condition (ibid).

An important clarification here is that they added another condition where they explicitly told the Asian-American participants to make trait attributions before seeing the face/behaviour pairs. After doing so, Asian-American accuracy levels on the incongruent traits lowered to equal that of the European-Americans (p. 1028). As such, we can be relatively secure in assuming that the study did track the effect of

(2008) argued that the presence of the P300 in relation to the observation of inconsistent trait behaviour showed that a previous spontaneous trait inference had been made.

[118] There are some questions as to what the cultural difference really was between these 'European-Americans' and 'Asian-Americans'. However, this was somewhat controlled for. Na and Kitayama originally predicted that individualistic vs. collectivist senses of self were in-part responsible for cultural differences between these groups and hence for making spontaneous trait inferences. In study two, they gave the participants a questionnaire about their independent vs. interdependent senses of self, where they found that there *was* a difference between groups, and furthermore that this was partially mediated by self-construal according to regression analyses (p. 1030). As such, while we cannot really say what these differences were specifically, a construal of the self across an independent/interdependent spectrum, in a general sense, appears to reasonably track the underlying supposed cultural differences between European-Americans and Asian-Americans in this study. For additional corroboratory evidence across distinct geographic-cultural boundaries, see Miller (1984), Newman (1991), Choi et al. (1999), and Kuwabara et al. (2011).

spontaneous trait inferences, where this implies that previously inferred semantic information (trait attributions) were being considered (and then rejected or updated) in the European-American condition.

Study one, therefore, appeared to show that spontaneous trait inferences were being made for European-Americans, but not for Asian-Americans. This psychological effect can be tested further with neuroscientific data, hence the entrance of the N400 for study two. The presence of the N400 itself does not say much, because N400 appears to process semantic meaning. What is important for studies that detect relatively strong N400 activation (compared to control conditions) is that "the N400 is thought to index detection of semantic incongruity" (p. 1028). As such, the prediction is that if European-Americans had made spontaneous trait inferences in the memorisation trial, then a *strong* N400 response should be elicited when the incongruent trait is presented, where for Asian-Americans no such effect should be presented. It is in this way that this study encompasses a 'violation of expectation' paradigm.

The results of study two then confirmed this prediction that European-Americans were making spontaneous trait inferences, because they elicited strong N400 activation when incongruent traits were subsequently presented to them, whilst no such effect occurred for the Asian-American participants (p. 1029).

Therefore, whilst the N400 is itself a marker of semantic incongruity, as Na and Kitayama noted, "we predicted that the N400 would be a sensitive index of spontaneous trait inference, and we found that it was" (p. 1031).

Indeed, similar results can be achieved through cultural differences even within a wider cultural group. Varnum et al. (2012) noted that, using the same violation of expectation paradigm for the detection of spontaneous trait inferences through N400 activation, middle-class American participants made more spontaneous trait inferences than working-class participants (as defined—perhaps poorly—by the educational level of their parents).

As such, these data of cross-cultural variability in the neural indicators of spontaneous trait inference activations provides an empirical line of support for trait attribution as a gadget.

4.2.2 Objection: Cross-Cultural Variability in Automatic Processes do not Evidence Gadgetry

However, I will entertain a final quick objection. It has been suggested to me that in itself, cross-cultural variability in an automatic process does not evidence gadgetry. This objection is clarified with a case study in the perception of colour.

There is cultural variability in the perceptions of the categories of colour (e.g., between categories of blue and green) across languages—such differences can even impact processing time regarding perception of colour differences (Kay and Kempton 1984; Roberson and Davidoff 2000; Winawer et al. 2007; Thierry et al. 2009). Such categorisations are made purportedly automatically, but the critic might suggest that these facts do not, by themselves, evidence gadgetry. This is because the data shows that pre-verbal infants appear to possess a universal colour categorisation ability, such that they can categorise colours before they learn the words for blue and green, for example (Franklin, Skelton and Catchpole 2014; Witzel and Gegenfurtner 2018, pp. 486–498). The fact that this is *pre-verbal* is significant because colours lie on a spectrum—categorisations based on colour are therefore in some sense the result of biology or cognition.

As such, the objection is something like this: automatic cross-cultural variability does not evidence gadgetry because there is a non-gadget example (universally acquired pre-verbal colour categorisation) that nonetheless has a culturally variable automatic effect (in adults).[119]

Regarding this objection, I could concede the point and yet still note that my argument contains many moving parts; hence, my argument for the gadgetry thesis regarding trait attribution can remain more or less intact. This would be given the remaining outpacing genes argument, for example. Furthermore, I could note that even if this example demonstrates that cross-cultural variability in automatic processes is not, by itself, evidence for a gadgetry thesis, the fact that it is being packaged with other evidence for the gadget thesis *does* suggest that this is an apt matter of consideration. The claim, then, is that in isolation it may not be evidence but as part of a package it could raise the probability of the gadget thesis being correct. Indeed, I could even claim that the burden of proof is on the critic to show why this is not evidence against the opposing

[119] My thanks to Stephen Butterfill, for raising this point in conversation. However, note that my reconstructed representation of this argument might not be exactly as he intended.

view—whose account I detail in the following section—since the cross-cultural variability claim is complementary to the gadget thesis.

That said, I want to reply with two substantive comments. Firstly, one can query the extent to which adult colour categorisation is automatic. Recall my definition of automaticity from Chapter three. The issue was in whether the effect could be inhibited, and therefore some control exerted over the behaviour. If it could, then it was not automatic. Witzel and Gegenfurtner (2015) found that 'automatic' category effects on perception disappear when subjects are trained in fine colour discrimination, that is, the differences between colours themselves rather than colour categories. As such, unlike spontaneous trait inferences, the colour category effect is inhibitable. My claim about cross-cultural variability of automatic processes, then, can remain untarnished on my understanding of automaticity.

Secondly, I suggest that the claim that 'infant colour-category perception is not a gadget' is more contested than it first appears; the proposed counter example is therefore not as secure. If pre-verbal colour categorisation was *genetically* encoded, then we might expect some semblance of this evidenced in comparative psychology of evolutionary development. This does not seem to be the case. Baboons have similar colour vision to humans, however there is no evidence that they respond to colour categorisations, such as the difference between blue and green, or blue and purple (Franklin, Skelton and Catchpole 2014, p. 18). On a gadget account, on the other hand, this result would be explained by the differences between human and baboon brains in terms of which brains can build gadgets.

The final point regards the status of explanations for why pre-verbal infants are able to make colour categorisations. Franklin, Skelton, and Catchpole's (2014) meta-analysis of infant colour category perception presented several current theories (pp. 17-18). Of these theories, all seem to be compatible with learning to make these categories via domain-general associative learning mechanisms. Indeed, the most nativist theory discussed was the theory that these categories are hardwired into the visual system via the contrasts between colour cones (Xiao et al. 2011). However, Skelton et al. noted that this theory leaves open the question of how finer distinctions could be made, for while the cone theory works for red and green, it would not provide a distinction between red and yellow. Another theory mentioned that might be plausibly nativist was the suggestion that "adult colour categories could come from an interaction of perceptual inequalities in colour and general cognitive strategy to categorise" (Franklin, Skelton and Catchpole 2014, p. 17). However, in this case it is not clear why a disposition to categorise could not, in principle, be an element of the genetic starter kit upon which gadgets are built, and hence accommodated by the gadget account.

Furthermore, other current suggestions of accounts for this early-emerging ability are more obviously sympathetic to a thesis of domain-general learning through culture, than they are to a nativist thesis. For example, Witzel and Gegenfurtner (2018) suggest:

> Infants might acquire categorical information through shared attention and other kinds of interaction with their social agents (e.g., their parents) that do not depend on language. Cross-cultural commonalities and infant color categories could also have an ecological origin. For example, they could be related to statistical regularities of color distributions in the visual environment (…) infants might internalize these regularities early in development. In this way, the visual environment would shape color categorization through early experience rather than color categories being a consequence of hard-wired mechanisms of color processing (p. 490).

As such, due to the highly contested nature of accounts for pre-verbal colour categorisation, it is not clear to me that pre-verbal colour categorisation constitutes a counter example to the claim that cross-cultural variability in an automatic process evidences a gadget thesis. However, as I noted, even if this did constitute a counter example, I would still possess the theoretical resources necessary to make an argument for my claim that trait attribution is a cognitive gadget.

In summary for this section, I argued that trait attributions track features of the environment that change faster than genetic selection would allow for in selecting for such a mechanism, and that trait-attributive practices historically (and still do) vary across cultures. This might have evidenced only that the ability is merely learned, but I then argued that we can point to neural data on a cognitive mechanism of trait attribution that functions automatically but differently across cultures. However, though I entertained some objections to my positive claim, I have not yet given time to what I see as the most plausible competing account of the evolutionary underpinnings of trait attribution. Arguing against a strictly innate claim for the development of trait attribution is a standard affair, but the real challenge lies in rejecting what I call the 'thinnate' account; this forms the content of the final section of this chapter.

5. The Thinnate Account

Referring back to 2.2.2, I noted another relevant way in which the development of cognitive structures is effected. Such accounts appeal to 'thinly innately specified learning mechanisms': I call these 'thinnate' accounts. The relevant thinnate account discussed below is Peter Carruthers' account of mindreading (Carruthers 2013); pertinently, based on this work he has been involved in levelling thinnate objections against Heyes' claim that mindreading is a gadget (Roige and Carruthers 2019). First, I will outline the claims that a thinnate theorist makes, giving Carruthers' thinnate theory of mindreading as an example. Then, I will outline the evidence in the thinnate account's favour for trait attribution. I will then respond both that the gadget account can accommodate the new data that the thinnate account might appeal to, and that the gadget account is nonetheless better placed to explain such data.

5.1 How Mindreading as Thinnate Challenges the Gadget Thesis

On Heyes' view, domain-general associative learning mechanisms are responsible for most of our learning and acquisition of skills—the 'genetic starter kit' of innately specified information and mechanisms on which gadgets are built is rather minimal. Another sort of view that one might have is that evolution has specialised aspects of our cognition such that domain-*specific* learning mechanisms enable us to pick up specific skills. This is a nativist claim, as there are innately specified components, but it is weaker than a general innateness claim about a capacity because it only holds that *some* components of the skill are innately specified (hence 'thinly'), where the information innately encoded not only constitutes a basic capacity for the skill but is also structured to help one acquire the adult skill.

The thinly specified components of the learning system are "conceptual primitives and priors" (Roige and Carruthers 2019, p. 3) and "attribution rules" (ibid). Conceptual primitives and priors may be things like the ability to attribute the basic propositional representations involved in that capacity (Carruthers 2013, p. 143), whereas attribution rules are innately possessed elements of theory, such as 'if, then' statements about knowledge that can be gained from certain behaviours, for example that 'seeing is knowing' (Roige and Carruthers 2019, p. 7).

The motivation for adopting a thinnate view of any particular cognitive system pertains to one's stance on the evolution of cognition. On the strongly constructivist view, the flexible and powerful associative learning capacities that are demonstrable across the animal kingdom are the primary targets for evolutionary development, and enhanced cognition follows. However, another approach to the evolution of cognition is to compare it directly to biological evolution, given that the brain is, as a biological object, the product of natural selection in the same ways that bodies are. Much of evolutionary biology stresses the domain-specific selective add-ons that are built in and scaffolded on existing biological architecture during natural selection; the claim is that there are no reasons that the same would not apply to the mind (Carruthers, 2006).[120] As such, as adaptations are selected for and built in a domain-specific manner, we would expect some domain-specific learning mechanisms to work with domain-general associative learning mechanisms.

Such are thinnate theories generally, but relevantly Carruthers (2013) presents a thinnate theory of *mindreading*; he holds that adult mindreading has developed as a result of these thinly specified innate learning mechanisms: "there is an innately structured domain-specific learning system, which is designed to build the mature [mindreading] system in response to both direct experience and cultural input" (Roige and Carruthers 2019, p. 7).

There are some caveats to presenting this account. For example, Carruthers notes that a thinnate claim about mindreading "postulates an innately channelled body of core knowledge, or an innately structured processing mechanism (or both), with an internal structure that approximates a simple theory of mind" (Carruthers 2013, p. 151). There is some ambiguity here, considering that the innately channelled body of knowledge does not appear to be substantially different in kind from a set of conceptual primitives plus attribution rules, yet apparently this may come apart from a "structured processing mechanism" whilst achieving the same mindreading result. Furthermore, whilst Roige and Carruthers bill the thinnate account as a "learning mechanism", citing Carruthers (2013) as an example of this, it is not clear what the relevant learning mechanism looks like. Carruthers outlines the thinnate view of one where "the system is designed to enrich itself as development proceeds, acquiring new ways of inferring people's mental states from behavioural or contextual cues, for example" (p. 142). How might these new ways be acquired? Appealing to domain-general associative mechanisms, for example, does not seem compatible with the idea that the domain-specific system is designed to enrich itself, specifically. Furthermore, one may wonder how the capacity becomes

[120] One need not commit to a mind/brain identity thesis in order to admit that the structure of the brain affects how the mind operates.

enriched, given that in the course of development "no new mechanisms are built or come online. And no deep changes in the representational resources available for mindreading take place thereafter" (p. 167). That said, issues of clarity are not insurmountable. For example, maybe the thinnate theorist could appeal to the notion of special vocabulary, outlined in Chapter one, which clearly states how the structure of the theory itself guides the kinds of information that can be added to the theory. Furthermore, Carruthers' account appeals to interactions (though as yet undefined) between thinnate systems and planning and reasoning systems—perhaps much of the enrichment can occur through the way in which such interactions are structured.

Caveats aside, the way in which the thinnate account of mindreading constitutes a fair challenge to the gadget account goes beyond the motivations for a thinnate theory generally, given that such motivations bottom out in an unresolved disagreement about the nature of the evolution of cognition. The thinnate mindreading account, Carruthers claims, accommodates all of the data that Heyes appeals to, and particularly any cross-cultural variance. This is in how culturally distinct learning practices will affect, in different ways, which concepts and attribution rules get added to the system's knowledge-banks: "much of the input that the learning-system is designed to accommodate will comprise the diverse verbal practices of mentalizing description and explanation that develop within each culture" (Roige and Carruthers 2019, p. 7).

In demonstrating the case for how mindreading might be thinnate, we can see how the claim would apply to trait attribution. On such an account, one could make many of the same predictions about the developmental data (particularly cultural variance) as does the gadget account; hence, the thinnate account appears to be on explanatorily equal footing. The threat of this rival account is greater than this, though, because a thinnate theorist can appeal to data about the sorts of conceptual priors that a trait-attributive learning mechanism might have. I detail this below, before presenting a response in favour of the gadget thesis.

5.2 Interpreting Traits from Faces

Heyes notes that humans have a prenatal preference for faces. For example, in the womb, pre-natal infants respond to lights shone through the mother's stomach in the shape of a face (two lights for eyes and one for mouth), but do not respond when the shapes are upside down or disfigured (Reid et al. 2017).[121] An innate preference for faces, such that babies can be "extracting information as well as care from the adults around them" (2018, p. 53) from birth, constitutes part of the 'genetic starter kit' of pro-social biases upon which gadgets are built, according to Heyes.

However, the thinnate theorist may argue that the *trait* information extracted from faces by young children supports a thinnate thesis of trait attribution: whilst much of trait attribution is learned (and thus differs cross-culturally), very young children develop adult-like face-to-trait attribution skills (Cogsdill et al. 2014). In Cogsdill's study, firstly, adult participants viewed videos showing faces, and rated them high or low on traits related to those faces. When high/low variants of such traits were coded, the videos were then shown to children, who were asked questions like "which of these people is very nice?" (p. 1133). Crucially, the youngest participants were three years old, and were able to select the same results as the adults with a pass rate that was above chance (60-70%).

The claim that these children possessed adult-like fact-to-trait attribution is cashed out in terms of making judgments along trait-dimensions which were describable across 'warmth' and 'competence'. For example, stereotypically, old ladies are seen as warm and incompetent, whereas adult white males may be seen as cold but competent. The children in Cogsdill et al.'s study responded as adults do to facial features that were categorised by their differences in underlying warmth and competence dimensions, suggesting that very early onset and apparently adult-like functioning evidences some kind of innate learning mechanism for character reading—one that begins with dimensions of warmth and competence.

To clarify, supposing that the knowledge of character traits themselves is unlikely to be innate (c.f. Fodor 1975), the thinnate defender proposes that these data on warmth/competence support a claim that an innate learning system biased towards developing trait attribution is present from a very early age, guiding the trait-attributive capacity's development. As such, building on a small core body of knowledge (that allows for simple judgments which are describable along trait dimensions of warmth and competence) and influenced from learning, we have a thinly specified innate

[121] This study is fascinating, but the reader can be forgiven for being sceptical of the methodology. For replications and more detail on the significant preference for faces that infants display, see (Johnson et al. 1991; de Haan, Pascalis and Johnson 2002; Farroni et al. 2005; Johnson 2005).

(thinnate) thesis about trait attribution; this would constitute an empirically supported threat to my gadget thesis.

5.3 Responses to the Thinnate Data

My first response will be to show how the gadget account can accommodate the data. This is the defensive move; the subsequent offensive move will be to show that, currently, the thinnate account of face-to-trait attribution lacks an explanation of how one moves from understanding a face to attributing a trait, whilst a gadget account has an explanation at the ready. As such, the inference to the best explanation for the developmental underpinnings of trait attribution still lies with the gadget account.

5.3.1 Gadget Accounts can Accommodate the Evidence

Data on warmth and competence are compatible with Heyes' account of the 'genetic starter kit' upon which gadgets are supposedly built. The outpacing genes argument above noted that characteristic characterful behaviours changed so rapidly even within a homogenous community that trait-attributive capacities would not be selected for. That said, the dimensions of warmth and competence are clearly not full-blown character traits. For the sake of argument, suppose that there may be enough reason for mechanisms that track warmth and competence to be selected for. If this is the case, then such a bias could constitute part of the pro-social genetic starter kit upon which gadgets are built; as such, the predictions of the thinnate account and the gadget account would match one another. Indeed, this accommodation can be bolstered further by showing that a prediction of a gadget account about warmth and competence is borne out by the data.

Because extracting information from faces along dimensions of warmth and competence is an early-emerging cognitive ability, a gadget thesis would predict that some kind of ability to either extract data from faces or to recognise warmth/competence may be forthcoming in other animals. This is because the gadget thesis claims that instead of inheriting 'Big Special' psychological attributes (like ability for language), "we genetically inherit 'Small Ordinary' psychological attributes: the propensity to develop relatively simple mechanisms that closely resemble those found in other animals, including chimpanzees" (Heyes 2018, p. 53). It is by building upon the 'Small Ordinary' mechanisms that human-specific genetic starter kits have evolved; hence, it is from these that gadgets are eventually constructed. Given this, we would predict that there would be some resemblance in the recognition of warmth and competence among animals like chimpanzees, or some evidence that they extract data from faces.[122]

Whilst there is little evidence that chimpanzees extract information from others relating to warmth, there is evidence that they recognise competence. Melis et al. (2006) found that when given multiple partners for collaborative tasks, chimps recruited the more effective partners from prior tasks; chimps appear to be able to recognise the importance of competence, and to recognise competence in other chimps. This is not evidence of judgments that are made along implicit dimensions of warmth and competence, of course. What it demonstrates is that the competence concept can be instantiated in chimp minds. Not only can chimps recognise competence but they can also extract data from faces: Tomonaga et al.'s (2004) meta-analysis found that infant chimps recognised and were able to track their mother's face from being one month old, and recognise the direction of the gaze from faces at two months old (pp. 229-230).[123]

As such, a prediction of the gadget account regarding Small Ordinary psychological abilities has been borne out; hence, the warmth and competence data can be accommodated by the gadget account. One should note that this is a response that assumes a strong gadget thesis; that is, one that accepts the ubiquity-of-gadgets thesis and the human distinctiveness thesis. This is because the response, as given, takes Heyes' particular account of the genetic starter kit. As noted, I am not committed to this; I present this argument to demonstrate that even the strongest gadget thesis can resist the force of the thinnate claims, here.

5.3.2 Warmth/Competence is not Domain-specific to Trait Attribution

A gadget theorist can respond further, though, in claiming that warmth and competence do not appear to be domain-specific to trait attribution; it is unclear how a thinnate account for domain-specific learning mechanisms would square its theoretical

[122] It is unlikely that we would find both faculties in even our closest cousins, the chimps. The gadget thesis argues that humans have more powerful cognitive faculties than animals because our genetic starter kit is biased in a way that allows for powerful gadgets to be built. All animals have domain-general learning mechanisms (even snails, Acebes et al. 2012), but no animal even comes close to the human ability. As such, a gadget thesis would not predict that animals, like humans, would both be able to extract information from faces and be able to make judgments about warmth and competence.

[123] Comparatively, replicated studies found that human infants recognise their mothers from four days old (Pascalis et al. 1995).

commitments with the implementation of domain-general conceptual priors as the main force of the learning mechanism.

To elaborate, if trait attribution is thinnate, rather than learned or a gadget, then there should exist a core body of knowledge which is specific to the functioning of that capacity (rather than any domain-general body of knowledge which is accessible to domain-general learning systems). However, Ponsi et al. (2016) showed that within a second of the 'first sight' of a face, one appears to categorise the owners of these faces as in-group or out-group members per the dimensions of warmth and competence.[124] Indeed, dimensions of warmth and competence themselves stem from work on stereotype attribution (Fiske et al. 2002), and it has been claimed that warmth and competence are the 'universal dimensions of *social cognition'* (Fiske et al. 2007, emphasis mine). Whilst these different kinds of attributions are all socially cognitive, they are nonetheless distinguishable capacities. As such, the thinnate defender cannot claim that this body of knowledge is innate to a *trait*-learning capacity *per se*, though they may be able to make weaker claims about the understanding of others. For example, a weaker claim might be that the core body of knowledge enabling apparent warmth/competence judgments is actually part of a *social cognition* core body of knowledge.

This amended view is, however, unpalatable for objectors to the thesis that trait attribution is a gadget. Even if we grant that warmth/competence is domain-specific to social cognition, we circle back to the accommodation of data through the prediction of the same results: the gadget thesis holds that gadgets are built upon existing genetic bases, such that trait attribution stemming from social cognition systems is favourably comparable with the visual word form area stemming from object recognition systems.

In sum, the fact that warmth/competence judgments are not domain-specific to trait attribution raises questions for the thinnate thesis, but even if we suppose that what enables warmth/competence judgments is a domain-specific body of knowledge to social cognition, the thinnate thesis of trait attribution only has as much explanatory force as the opposing gadget thesis. I now move to where I believe that the gadget account has the advantage.

5.3.3 Thinnate Accounts Lack an Explanation of Moving from Face Interpretation to Trait Attribution

[124] 'In-group' and 'out-group' being terms used in social identity theory (Tajfel and Turner 2010).

Going on the offensive, the gadget theorist can note that a thinnate theorist has (as of yet) no unified account of the development of trait attribution derived from face recognition. However, there does exist such an account that a gadget theorist of trait attribution may use. This is the 'trait inference mapping' account of Over and Cook (2018). Building on existing computational accounts of the way in which we recognise faces by mappings in a 'face space' (Valentine and Ferrara 1991; Valentine, Lewis and Hills 2016), their account introduces the 'trait space' as a similar encoding of vectors, such that mappings "allow excitation [activation of stimuli] to spread automatically from face representations to trait representations, and thereby give rise to spontaneous trait inferences" (Over and Cook, p. 191).

Crucially, face-trait mappings are products of domain-general associative learning processes (p. 197), where the development of trait knowledge and mappings of the vectors between the face space and trait space is a product of culturally inherited information. For example, such mappings of vectors can be inherited from cultural artefacts like the faces and perceived characters of figures in religious iconography, the traits and appearances of characters in fairytales, and the modern TV and film depictions of heroic or evil figures, to name a few (p. 194).

Whilst being detailed and empirically supported in its own right, its mere existence as a unified account of the development of trait attribution based on faces is a point in favour of the thesis that trait attribution is a gadget. This is supposing, of course, that the thinnate theorist cannot also appeal to these data. Can they? I think not—at least, not in a way that matters for the gadget theorist. Consider the following: suppose that we gave some thinnate account of trait attribution—such an account may be conceptualised in two ways: either trait attribution is itself thinnate (hence, there is a thinly innately specified domain-specific learning system for attributing traits), or trait attribution is an add-on to the thinnate capacity for mindreading.

On the former view, where there is an innate capacity for minimal trait attribution that helps the acquisition of the more complex adult skill, such an account cannot help itself to the explanatory power of the Over and Cook account. This is because an explanation of how trait attribution functions on a specifically thinnate understanding cannot appeal to domain-general learning mechanisms for the attribution itself, i.e., for the basic functioning of the innate component; domain-general mechanisms may only aid in the addition of new concepts, priors, and attribution rules. Perhaps a thinnate theorist might respond that there is a minimal innate capacity for attributing traits, but attributing traits *on the basis of facial displays* is bootstrapped from some additional primitives and priors generated by the mappings from the face space to the trait space, which is allowed for on a thinnate account. This seems like a reasonable response at first, but then one must wonder what prior features there are that a thinnate trait attribution system could use to attribute traits—faces and their emotions seem like the most basic application of moving from emotion attributions to trait attributions. If the thinnate theorist attempts to accommodate Over and Cook's account along these lines, then, this is tantamount to admitting that the innately specified components of the trait attribution system are non-functional in the most basic instances; the thinnate account would be watered down to have basically no explanatory power.

On the latter view, the thinnate component is the mindreading mechanism, and the trait-attributive capabilities are an add-on, presumably built through new concepts, priors, and attribution rules for the system. This style of thinnate account of trait attribution could indeed appeal to the same domain-general learning data that Over and Cook use, but at a cost. If trait attribution is an add-on, then it would count under the extra acquired conceptual resources that then interact with planning systems and reasoning systems (on Carruthers' construal of the thinnate account of mindreading). However, as noted above, the thinnate account supposes that core representational structures are basically already in place (Carruthers 2013, p. 151), such that there should be no development of the mindreading mechanism *per se*. If this is the case, then we should not expect something like trait attribution to become situationally autonomised through an *additionally* developing cognitive process that is—in principle—not accessible to consciousness, i.e., spontaneous trait inferences. But let us suppose that the thinnate theorist can produce some explanation here, some more accommodation. Even then, considering trait attribution to be part of the mindreading architecture raises the following question: What is the functional difference between thinnate trait attribution as a learned reasoning skill built on top of mindreading, and it being a cognitive gadget built on top of an innate mindreading structure? It seems that cutting the thinnate claim this way reduces the debate to one of semantics, rather than a substantive disagreement about cognitive architecture.[125]

[125] One might suppose that the debate instead bottoms out in the nature of the evolution of cognition, as noted above, but this is irrelevant for a gadget thesis, given that a thinnate account can allow that cognitive gadgets exist (Roige and Carruthers 2019, p. 2).

As such, the thinnate theorist cannot appeal to Over and Cook's account to plug the explanatory gap, but what other resources might they have to explain the move from facial processing to trait attributions? Perhaps there is some view that I have left unaccounted, but it seems to me that they (currently) only have arguments from standard evolutionary psychology regarding the potential benefits of specific face-to-trait attributions, such as recognising trustworthiness for identifying friends and foe (Todorov et al. 2008), or dominance to identify leaders (Vugt and Grabo 2015). However, even if these were granted as selectively beneficial, I have already argued in 4.1 that trait attributions track labile features of the environment, so such specific traits are unlikely to be selected for; it is hard to justify why such face-trait mappings would be encoded into the genes.

In summary, this section has presented the most plausible alternative account of the developmental underpinnings of trait attribution. I argued that, in the specific case of trait attribution, the gadget account and the rival 'thinnate' account both make successful theoretical predictions that are borne out by the same data; hence, they are explanatorily equal on that front. I then showed that a significant set of data that purportedly favours the thinnate account has an explanatory gap that the gadget account already fills, in explaining how one moves from seeing faces to attributing traits—I then showed why the thinnate account cannot accommodate that data. As such, I have defended my claim that trait attribution being a gadget is a better explanation of the existing data.

Conclusion

This chapter outlined and defended my thesis that trait attribution is a cognitive gadget, that is, a psychological capacity of which the cognitive mechanism is the result of cultural, rather than genetic, evolution. This thesis follows from the previous chapter's arguments that character reading is conceptually and explanatorily dependent on mindreading—I supposed that some sort of ontological dependence for the capacity existed between the two that may further explain this relationship; the gadget thesis neatly provides this through its account of the construction of cognitive mechanisms based on existing cognitive machinery (such as mindreading mechanisms).

I gave my positive case for trait attribution being a gadget with a theoretical argument and an empirical one. Firstly, I argued that trait attribution tracks labile features of the environment that move too quickly for natural selection to encode in genes. Secondly, I argued that a cognitive mechanism for trait attribution, the spontaneous trait inference mechanism, was a cognitive gadget. This was on the basis that the automatic attributions of character made by the mechanism varied cross-culturally. After presenting my positive case, I considered a plausible nativist alternative for the development of trait attribution, in the form of the thinnate account, such that trait attribution would be the product of a genetically encoded domain-specific learning mechanism that helps us to enrich our trait-attributive capacities as we take in relevant cultural information. I argued that the thinnate account and the gadget account both accommodate the data, but the difference maker in favour of the gadget account is demonstrated when we consider how one might move from recognising faces to attributing traits.

Finally, a note about culture. In the introductory chapter of this thesis, I endorsed Jane Suilin Lavelle's (2021) claims that good socially cognitive accounts need to provide the conceptual resources to explain how culture affects our understanding of others' psychological states, and that they need to show that differences in culture are not irrelevant to the cognitive architectures of our socially cognitive capacities. I promised to make good on this goal in this chapter. I take it that, in recognising trait attribution as a gadget, we have the conceptual resources (that go beyond an endorsement of folk-psychological pluralism) to explain the cultural differences between practical applications of our understandings of each other's characters; furthermore, we can account for the impact of culture on cognitive architecture through the culturally evolutionary underpinnings of our trait-attributive abilities.

Character and Culture in Social Cognition

1. Answers to the Thesis Questions

In writing this dissertation, my hope is that across the interdisciplinary topics I have covered, I have illuminated some issues of note and answered questions about others. There are, of course, specific questions that this dissertation sought to answer. After a reminder of these questions, I will move to a discussion of what, precisely, my contributions to this discipline have been. This will also serve as an opportunity to discuss why the structure of this dissertation is as it is: why I set up the dialectic in the ways that I did.

I posed several research questions in the introductory chapter. For example, I asked 'how should we understand the ontology of character traits?' By this, I meant not only the metaphysics, but the cognitive architecture and processing of traits in social cognition. The answers should be clear, now. Metaphysically, character traits are dispositional abstract entities: psychological tendencies that consist of dispositions towards behaviours and psychological states, coupled with implicit histories or summaries of past disposition manifestations. In terms of cognitive architecture, the capacity for character trait attribution is a cognitive gadget, that is, its cognitive mechanisms are the products of evolution acting on culture, where character trait attributive architecture has been constructed upon existing mindreading architecture. The processing of character traits in social cognition is achieved through a hybrid account of the theoretical and simulative processes that we engage in.

Another question that I posed in the introduction asked how our ability to attribute traits to others functions when understanding others through their character? Whilst heavily weighted towards general theoretical inferences based on one's tacitly held theory of traits, understanding others' character can be simulative when we either generate imaginations from episodic memory to make trait attributions, or when we holistically consider the practice of character reading when it involves simulative practical reasoning.

I also posed a third question: what is the relationship between character reading and mindreading? The answer to this question is detailed through my articulation of the dependence relation, such that the capacity for character trait attribution depends on mindreading in a conceptual and psychologically explanatory sense. This, however, does not mean that character trait attribution requires mental-state attribution, and it does not mean that the capacities are one and the same.

In addition to posing these general questions, in the introductory chapter I also made clear the claims that I was building towards in writing this dissertation. Those were the following:

- Character reading is a socially cognitive skill that differs from 'mindreading' on theoretical, metaphysical, and empirical grounds.
- The cognitive processing involved (both conscious and unconscious) in how we character read can be captured by 'hybrid theory/simulation' accounts.
- Due to the situating of character reading within an emerging 'pluralistic' understanding of socially cognitive skills, of seeing how character reading happens unconsciously and consciously, and of understanding how trait attribution depends on a 'mindreading' skill, trait attribution should be considered a 'cognitive gadget', a cognitive mechanism that is the product of cultural, rather than genetic, evolution.

I take it that the above claims have been made and defended over the course of this dissertation. They are, however, not obviously connected to each other in a cohesive narrative when considered out of context like this. The narrative is apparent across the chapters, but what remains now are the tasks of making clear the value of the dissertation and the theses it argues for, by answering the following questions: What, specifically, were the original contributions made by this dissertation, and how does it affect current scholarship on these issues? What were the limitations of my approach? What is the appropriate direction of future research implied by this work? I turn to this discussion now.

2. The Contributions of the Research

To ascertain what the contributions were in this dissertation, we need to understand what my objectives were in my writing of it. I wanted to contribute to ongoing conversations in the philosophy of psychology, particularly in my area of specialisation: social cognition. To that end, I saw folk-psychological pluralism as an exciting new direction for research that was sensitive to cultural variability in psychology. This thought grew out of the first piece of reading that I did in my first week of reading for this doctorate: Shannon Spaulding's (2018) book '*How We Understand Others: Philosophy and Social Cognition'*. Her discussion of 'model theory' opened my eyes to different ways of conceiving of the literature on mindreading beyond the work on false-belief tasks that I was familiar with. This research eventually led me to Kristin Andrews' (2008) paper on pluralistic social cognition, in which she uses character trait attribution as the paradigmatic example of a pluralistic folk-psychological practice. I was taken in by how obviously true that the thesis seemed—of course we understand others in a variety of ways beyond merely processing their presumed beliefs and desires. Of course, these pluralistic practices will differ across cultures. I sought, therefore, to help clarify the folk-psychological pluralist position by contributing to research on their paradigm example: character trait attribution. I believe that I have made a contribution towards understanding how this ability functions and how the research program of pluralism is complementary to existing scholarship on social cognition.

I noted that a consideration of the pluralistic practices across cultures was important to consider. Indeed, I see psychological work as having been so focussed on what sorts of universal psychological claims that may be made that, until recently, consideration of the variations between cultures across the world have been lacking. This is why the introductory chapter makes a clear commitment to Suilin Lavelle's (2021) imperative that good socially cognitive accounts need to clearly show where culture makes a difference (p. 6352). Fortunately, the literature on the impact of culture has been developing over the last few years. Evolutionary psychology appears to be pivoting towards research into how evolution and culture combine—it is fortunate that I was able to show the impact of cultural evolution on character trait attribution. I take my aim of increasing the salience of cultural impact in social cognition to have been achieved, however marginally, by the claims that I argued for in this dissertation. I now turn to a chapter-by-chapter summary of my research motivations and the contributions that each chapter has made—I discuss limitations and paths for further research along the way.

2.1 Chapter Contributions

In the introductory chapter, I demonstrated how forty years of psychological and philosophical research on social cognition has culminated in a conversation about the nature of socially cognitive practices that are seemingly distinct from, yet related to, mindreading. Character trait attribution is an example, but using stereotypes and following scripts and schemas are others. If the folk-psychological pluralist movement is to make any waves in helping us to understand our suite of varying methods for understanding others, then such a relationship needs to be made clear with a thorough account of a pluralistic practice: I chose character trait attribution, given Andrews' lengthy treatment and defence of it against the primacy of mindreading.

In Chapter one, I presented my 'tidying up' of the literature, particularly in the disambiguation of several senses of mindreading. Often it is unclear whether authors are appealing directly to a process of mental-state attribution, or whether they are treating mindreading in a folk-psychological sense of a cluster of skills that we use to understand others' minds. Indeed, folk psychology is also a term with many similarly (yet intricately different) uses in the literature, which I attempted to make clear for those interested in this topic. The choice in this dissertation to treat mindreading as specifically 'mental-state attribution' is merely a choice, but the value can be found in both the articulation of the choice, and my use of the term to show how mindreading relates to character reading in further chapters. Furthermore, I showed how on a particular understanding of what a theory is, theory and simulation are essentially the only two games in town for understanding the methods of socially cognitive understanding. This cleans up the dialectic, whilst retaining space for the 'alternative' conceptions that have often been suggested. On my view, theory-theory accounts are highly varied in structure, as it is the special vocabulary that gives the theory its structure. As such, all radical variants or attempts to occupy new areas in the logical geography are mistaken—they are likely just accounts that employ interesting kinds of theory.

This wide variation in theoretical structure on one side of the equation suggests that simulation, too, might be varied in its application. Indeed, whilst this dissertation almost exclusively dealt with Goldman's notion of process simulation, there are of course Gordon's (1986) account of simulation and my notion of experienced simulation, which was developed from Jane Heal's (1998) work on co-cognition; the banner of simulation can therefore be as wide as theory. Indeed, I make this explicit in Chapter four, when I argue for simulation in places that people are not usually looking. As such, whilst Chapter one serves to define my terms, I hope to have also contributed to how we think about this literature. The takeaways are, firstly, that theoretical inference and simulation are the most apt ways to describe the operation of our socially cognitive processes. Secondly, we need to be careful and provide clarity about the precise nature of terms like 'mindreading' or 'folk psychology' when we use them, given the history of crosstalk with such terms.

Whilst Chapter one defined terms, there was not the space to define the actual target phenomenon of this dissertation: character traits. In Chapter two, I detailed what I take a character trait to be, and defended my definition from both metaphysical and practical objections. I made it clear that whilst a dispositional sense of character traits is apt, character traits are not mere dispositions. What are crucial to what character traits are, are the histories of disposition manifestation that accompany the dispositions. These histories of manifestation not only make traits what they are, from a metaphysical perspective, but also distinguish them from the attribution of emotions or other fleeting states.

As such, I championed the term 'tendency' for character traits. I showed that despite some hesitancy by Maria Alvarez to commit to the term (2017, pp. 85–86), it serves a suitable purpose. This purpose is in-keeping with not only the metaphysics, but also the ways in which authors refer to character traits in literature, and how folk conceptualise traits. This is important, because putting forward a definition of character traits that did not match up with folk expectations would have been disastrous, given that many of the cited empirical studies in the dissertation ask the participants to consider character traits and yet let them rely on their folk interpretations as to what that means. Finally, I made sure that in the giving of my definition, I was as neutral on the precise metaphysics as possible, for the acceptance of my thesis on character traits in social cognition should not be dependent on one's personal metaphysics.

In Chapter three, I attempted to begin understanding how the process of character trait attribution might function. Given some acknowledged, but as of then undefined, relationship between character trait attribution and mindreading, I sought to compare and contrast mindreading accounts of such functioning against character trait attribution. Of the approaches I could have taken here, I chose to investigate whether character trait attribution could be achieved purely simulatively. This is because the (minimal) existing scholarship on the issue unanimously asserted that it was philosophically implausible. This presented an interesting challenge to me, for I wanted to see if simulation in trait attribution *could* be made plausible, given that no thorough investigation on the topic had been performed. My choice to research Alvin Goldman's (2006) work on simulation, for such a challenge, was because it was the most comprehensive and detailed work on simulation available.

Of course, the outcome of Chapter three is that a simulationist theory of traits in social cognition is indeed philosophically implausible; however, now the due diligence has been performed on testing that claim. This alone is a contribution to the literature, albeit it a small one. That said, I supplemented the philosophical implausibility of the simulationist claim by consideration of much empirical data on the subject—it is through the discussion of Goldman's work that we can rule out pure simulation in trait attribution on both empirical *and* philosophical grounds. Whilst an overall negative conclusion to the chapter, what was promising was the fact that a well-detailed account of simulation in mindreading could not simply be transposed on to an account of simulation in character trait attribution. This gave the folk-psychological pluralist more fuel for their claims that such capacities are relevantly distinct from mindreading—if they were not so distinct, we would have expected the transposing task to be achievable.

In Chapter four, I sought to provide a proper account of the functioning of not only character trait attribution, but also the socially cognitive reasoning that is involved regarding character traits. Firstly, the contribution of this chapter was that I made the role clear for simulation in character trait attribution. This was made possible because the investigation in Chapter three concluded that whilst a simulationist theory about traits was implausible, there was still room for a hybrid theory. In particular, there was space for some simulation involved in simulating the particular markers of personality that could be imitated. However, without much empirical data on the subject, I did not feel confident in making further claims in that regard. The possibility gave me the idea, though, that simulation in inaccessible processing was likely to be done with neural reuse (if it was achieved at all). Inspired by recent philosophical discussions of mental time travel, I noted that on an account where episodes of mental time travel are essentially imaginations that are reuses of memory architecture, then when one uses mental time travel to make a trait attribution, such simulations can result in trait attributions. The challenge I had set in undertaking Chapter three had been met in a technical sense, then, but of course the mere attribution is not all that there is to character reading.

In the second half of Chapter four, I tried to find ways to justify my intuition that there could be, sometimes, *something* simulative about the processes of understanding others by their character. It was not just this intuition that spurred me on, but also that I take the state of the debates over theory-theory and simulation theory to be essentially settled—in so far as hybrid accounts are the correct kinds of account. I thought that it would be highly unusual, therefore, for character reading to require a theory-theory account. I found the sense of simulation that I was looking for when distracting myself with television.

I claimed that Jake Peralta really did simulate the character of Amy Santiago, but I still needed to articulate this precisely. I realised that regardless of anything else, Westra (2018), and others, were likely correct that traits simply are not the sorts of things that can be inputs to practical reasoning simulations. I was at a loss for how to articulate my claim of simulation until I realised that I was essentially playing by their rules. Why does simulation need to be precisely where the theory theorist looks but cannot find it?

Indeed, I think that a contribution of this chapter to the literature generally is that whilst Goldman's focus on simulative processes in particular areas of cognition are important, the phenomenon of simulation occurs in many places and across many domains—indeed I now think that theory and simulation are absolutely crucial for understanding the operation of cognition across many levels of description. The holistic consideration for simulation in character reading is probably the least analytical (in style of) claim that I made, so I attempted to bring the boundaries of my claim into focus by relating the phenomenon to the technical definition of simulation that Goldman gave.

As such, my contribution to an account of character reading in social cognition was that, yes, a hybrid theory is apt. This is because simulation can sometimes, contra all other authors of the topic, be involved. What remains to be seen, though, is the precise special vocabulary associated with our theories of traits, where such would allow us to engage in practical reasoning about traits. A direct implication of my work thus far is that there must be such a plausible account to be given, so this is placed on my list of 'further research to be done'.

However, I am not convinced that such an account can be given any time soon in a comprehensive manner. For example, in Chapter six when I discussed the Over and Cook (2018) account of how we might move from seeing facial displays to attributing traits, such an account would clearly be only *a part* of the picture for an overall theory that relates situations and behaviours to traits. Indeed, such a theory might even need to be coherently related to other relevant theories in pluralistic folk-psychology. In particular, we might ask: what is the relationship between character traits and stereotypes? It seems that a theory of traits might need to include stereotypes, given that stereotypes often invoke character traits. This, unfortunately, is a clear line of research that I did not have the time for; hence, it is an important issue of further research. Nevertheless, I am optimistic on the project of detailing all of the elements of a theory of traits, given the allowance of the inclusion of mental-state information in such theories that I argued for in Chapter five; such inclusion seemed to be the missing link that Evan Westra noted of the folk-psychological pluralist.

In Chapter five, I wanted to make clear the boundaries between mindreading and character reading, by detailing what I saw their relationship to be. There is certainly a question of what this relationship is, given that Kristin Andrews takes trait attribution to be completely distinct from mindreading, whereas Evan Westra takes trait attribution to essentially be part of mindreading. My own investigation on the matter settles it somewhere in between, in a manner that I do not think either author has yet considered.

The issue is in the balance. Andrews is right that character trait attribution is a distinct skill from mindreading, if only because we can attribute traits without attributing mental states. She is wrong, however, that there is no substantive relationship between the two skills. Westra, on the other hand, is right that there is a clear relationship to be explained and he is right to be a mindreading primacist—it does seem as if our capacities for understanding others are crucially couched in our abilities to attribute mental states to them.[126] Westra is wrong, however, that the folk-psychological pluralist position on traits can be rejected *because of* mindreading primacy.

The line I trod between the two accounts was made possible by my understanding that just because character trait attribution might depend, in certain key ways, upon mindreading, this did not mean that mental-state attribution was required. Neither does it mean that character trait attribution is part of mindreading, and neither does it mean that the primacy of mindreading need be rejected. I made this clear by the disambiguation of Andrews' shorthand claim objection from the dependence claim objection. I then gave as much detail as I could about what this dependence relationship entails. I argued that there is a conceptual and explanatory dependence between the two where this is enough, I claimed, to tread the fine line between Andrews' and Westra's positions. This was especially demonstrated in my demonstration that Andrews'

[126] This endorsement of mindreading primacy also fits my theoretical commitments of not throwing out the last forty years of scientific research on social cognition.

evidence for the distinctiveness of character trait attribution from mindreading does not hold up to scrutiny.[127]

As such, we were left at the end of Chapter five knowing that in the social cognition of character, such accounts needed to be hybrid accounts. Furthermore, one could be a folk-psychological pluralist and yet still be conceptually committed to the primacy of mindreading. This I take to hold at least for character traits, but I also see it as a proof of concept for generally holding both views simultaneously. What remained to be understood, though, was the ontological relationship between mindreading and character reading. Furthermore, how the pluralist's commitment to culturally sensitive psychology affects our understanding of character reading—i.e., how I would meet Suilin Lavelle's cultural imperative regarding socially cognitive accounts.

In Chapter six, I sought to make good on those issues by detailing what I saw as the developmental origins of our trait-attributive capacities. Doing so was true to my roots in studying the developmental psychology of mindreading during my MA, but it was also well-timed as I was able to include discussion of Cecelia Heyes' account of the development of our socially cognitive capacities—the account is highly culturally sensitive; so much so that it is front and centre to her account.

Such would be interesting, but irrelevant, if character trait attribution was not a gadget. However, in my research on the empirical data regarding character traits, and in my philosophical reflections on such abilities over the course of reading for my dissertation, I came to realise that a case for the capacity being a gadget could be made. Furthermore, it could be made regardless of whether mindreading itself was a gadget.

Such a claim has not been articulated or argued for before, but in so doing I also contributed to some extra tidying-up of the literature. This was so by making it clear that in debates over the innateness of certain psychological capacities, the correct target of investigation are those 'thinnate' accounts that constitute a plausible (both empirical and philosophical) challenge to an empiricist thesis; nativism is essentially a strawman.[128]

[127] Westra agrees that Andrews' Social Stories™ evidence is lacking, and he has said so in print: Westra (2021, pp. 8217–8218). In an unfortunate coincidence, this was published at the same time as I was doing that particular research. That said, my contribution is apparent in the extra detail that I have gone to in assessing the data (with the meta-analysis reviews), and in my subsequent arguments about the scepticism regarding the possession of these concepts by the children.

[128] Though, notably, Stephen Laurence and Eric Margolis have been arguing for a version of this claim since 2001 (Laurence and Margolis 2001).

As such, Chapter six not only continued to detail the relationship between character reading and mindreading, it helped justify (by giving another relevant example) a new and exciting thesis about the development of cognitive machinery, that of the gadget thesis. The chapter also contributed to the debates surrounding nativism vs. empiricism about socially cognitive capacities, and it demonstrated clearly where the cultural impact is regarding character trait attribution. I take it that the cultural-evolutionary underpinnings of the capacity are more than enough to determine how and where the impact of culture affects the psychological capacity.

I had two frustrations with this chapter, though. Firstly, there was a lack of empirical work on character traits that was pertinent to the discussion, and secondly, it is not yet clear what the developmental timing is between mindreading and character reading. I was expecting to find research that confirmed that character trait attribution and reasoning developed after mindreading competency, but the current data suggests that they develop at around the same time. As noted in the chapter, our methodologies are not comprehensive enough for any more specificity than this 'around the same time' claim. As such, further research needs to be done in order to definitively rule out the claim that character trait attribution is essentially a form of mindreading; lots of this needs to be empirical (rather than philosophical) work.

In summary, whilst Chapters one, two and three were scene-setting (though they nonetheless each make scholarly contributions in their own rights), my account of the social cognition of character was given across the final three chapters of the dissertation. I argued for a folk-psychological pluralist view of character trait attribution, but one that is nonetheless sympathetic to the primacy of mindreading. I also detailed the clear relationship of dependence that holds between mindreading and character reading. This situates the skill within an emerging pluralistic understanding of our folk-psychological practices, without invalidating existing work on mindreading. I argued that such a theory must be a hybrid theory because there are instances where simulation is involved (both in the attribution and the reasoning about character). As such, hybrid theories of social cognition continue to be the most explanatorily apt. Finally, I argued that character trait attribution is a cognitive gadget, and so contributed to new and exciting scholarship on cultural evolutionary psychology. As such, I have produced a dissertation that argues for the precise nature of *character and culture in social cognition*.

Bibliography

Acebes, F. *et al.* (2012) 'Associative learning phenomena in the snail (Helix aspersa): Conditioned inhibition', *Learning & Behavior*, 40(1), pp. 34–41. Available at: https://doi.org/10.3758/s13420-011-0042-6.

Addis, D.R. *et al.* (2009) 'Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering', *Neuropsychologia*, 47(11), pp. 2222–2238. Available at: https://doi.org/10.1016/j.neuropsychologia.2008.10.026.

Addis, D.R., Wong, A.T. and Schacter, D.L. (2007) 'Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration', *Neuropsychologia*, 45(7), pp. 1363–1377. Available at: https://doi.org/10.1016/j.neuropsychologia.2006.10.016.

Adolphs, R. *et al.* (2000) 'A Role for Somatosensory Cortices in the Visual Recognition of Emotion as Revealed by Three-Dimensional Lesion Mapping', *The Journal of Neuroscience*, 20(7), pp. 2683–2690. Available at: https://doi.org/10.1523/JNEUROSCI.20-07-02683.2000.

Ahluwalia, A. (1978) 'An intra-cultural investigation of susceptibility to "perspective" and "non-perspective" spatial illusions', *British Journal of Psychology*, 69(2), pp. 233–241. Available at: https://doi.org/10.1111/j.2044-8295.1978.tb01653.x.

Ali, S. and Frederickson, N. (2006) 'Investigating the Evidence Base of Social Stories', *Educational Psychology in Practice*, 22(4), pp. 355–377. Available at: https://doi.org/10.1080/02667360600999500.

Allport, G.W. (1931) 'What is a trait of personality?', *The Journal of Abnormal and Social Psychology*, 25(4), pp. 368–372.

Altick, R.D. (1979) *Lives and letters: a history of literary biography in England and America*. Westport, Conn: Greenwood Press.

Alvarez, M. (2017) 'Are Character Traits Dispositions?', *Royal Institute of Philosophy Supplement*, 80, pp. 69–86. Available at: https://doi.org/10.1017/S1358246117000029.

Anderson, M.L. (2014) *After Phrenology: Neural reuse and the interactive brain*. MIT Press.

Andrews, K. (2008) 'It's in your nature: a pluralistic folk psychology', *Synthese*, 165(1), pp. 13–29. Available at: https://doi.org/10.1007/s11229-007-9230-5.

Andrews, K. (2012a) *Do Apes Read Minds?: Toward a New Folk Psychology*. Cambridge, Mass: MIT Press.

Andrews, K. (2012b) *Do Apes Read Minds?: Toward a New Folk Psychology*. Cambridge, Mass: MIT Press.

Andrews, K., Spaulding, S. and Westra, E. (2021) 'Introduction to Folk Psychology: Pluralistic Approaches', *Synthese*, 199(1), pp. 1685–1700. Available at: https://doi.org/10.1007/s11229-020-02837-3.

Apperly, I.A. (2008) 'Beyond Simulation–Theory and Theory–Theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind"', *Cognition*, 107(1), pp. 266–283. Available at: https://doi.org/10.1016/j.cognition.2007.07.019.

Apperly, I.A. and Butterfill, S.A. (2009) 'Do humans have two systems to track beliefs and belief-like states?', *Psychological Review*, 116(4), pp. 953–970. Available at: https://doi.org/10.1037/a0016923.

Aristotle (2019) *Nicomachean Ethics*. 3rd edn. Hackett Publishing.

Ashton, M.C. (2013) *Individual differences and personality*. Edition 2. Amsterdam ; Boston: Academic Press is an imprint of Elsevier.

Ashton, M.C., Lee, K. and de Vries, R.E. (2014) 'The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: a review of research and theory', *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, 18(2), pp. 139–152. Available at: https://doi.org/10.1177/1088868314523838.

Babel, M. (2012) 'Evidence for phonetic and social selectivity in spontaneous phonetic imitation', *Journal of Phonetics*, 40(1), pp. 177–189. Available at: https://doi.org/10.1016/j.wocn.2011.09.001.

Baillargeon, R. (1987) 'Object permanence in 3½- and 4½-month-old infants', *Developmental Psychology*, 23(5), pp. 655–664. Available at: https://doi.org/10.1037/0012-1649.23.5.655.

Baillargeon, R. *et al.* (2015) 'Psychological and sociomoral reasoning in infancy.', in M. Mikulincer et al. (eds) *APA handbook of personality and social psychology, Volume 1: Attitudes and social cognition.* Washington: American Psychological Association, pp. 79–150. Available at: https://doi.org/10.1037/14341-003.

Baillargeon, R., Scott, R.M. and Bian, L. (2016) 'Psychological Reasoning in Infancy', *Annual Review of Psychology*, 67(1), pp. 159–186. Available at: https://doi.org/10.1146/annurev-psych-010213-115033.

Baillargeon, R., Spelke, E.S. and Wasserman, S. (1985) 'Object permanence in five-month-old infants', *Cognition*, 20(3), pp. 191–208. Available at: https://doi.org/10.1016/0010-0277(85)90008-3.

Baron-Cohen, S., Leslie, A.M. and Frith, U. (1985) 'Does the autistic child have a "theory of mind" ?', *Cognition*, 21(1), pp. 37–46. Available at: https://doi.org/10.1016/0010-0277(85)90022-8.

Baumeister, R.F. (1987) 'How the self became a problem: A psychological review of historical research', *Journal of Personality and Social Psychology*, 52(1), pp. 163–176. Available at: https://doi.org/10.1037/0022-3514.52.1.163.

Bayne, T. *et al.* (2019) 'What is cognition?', *Current Biology*, 29(13), pp. R608–R615. Available at: https://doi.org/10.1016/j.cub.2019.05.044.

Beling, I. (1929) 'Über das Zeitgedächtnis der Bienen [Further investigations on the temporal memory of bees]', *Zeitschrift für vergleichende Physiologie*, 9(2), pp. 259–338. Available at: https://doi.org/10.1007/BF00340159.

Bem, D.J. (2011) 'Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect.', *Journal of Personality and Social Psychology*, 100(3), pp. 407–425. Available at: https://doi.org/10.1037/a0021524.

Bennett, J. (1978) 'Some remarks about concepts', *Behavioral and Brain Sciences*, 1(4), pp. 557–560. Available at: https://doi.org/10.1017/S0140525X00076573.

Berger, J.O. (1988) *Statistical decision theory and bayesian analysis*. 2nd ed. New York: Springer (Springer series in statistics).

Bisiach, E. and Luzzatti, C. (1978) 'Unilateral neglect of representational space', *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 14(1), pp. 129–133. Available at: https://doi.org/10.1016/s0010-9452(78)80016-1.

Blakemore, S.-J. *et al.* (2005) 'Somatosensory activations during the observation of touch and a case of vision–touch synaesthesia', *Brain*, 128(7), pp. 1571–1583. Available at: https://doi.org/10.1093/brain/awh500.

Block, N. (2005) 'Review of Alva Noe, Action in Perception', *Journal of Philosophy*, 102, pp. 259–272.

Bloom, P. and German, T.P. (2000) 'Two reasons to abandon the false belief task as a test of theory of mind', *Cognition*, 77(1), pp. B25–B31. Available at: https://doi.org/10.1016/S0010-0277(00)00096-2.

Bohl, V. and Gangopadhyay, N. (2014) 'Theory of mind and the unobservability of other minds', *Philosophical Explorations*, 17(2), pp. 203–222. Available at: https://doi.org/10.1080/13869795.2013.821515.

Bommarito, P.A. and Fry, R.C. (2019) 'Chapter 2-1 - The Role of DNA Methylation in Gene Regulation', in S.D. McCullough and D.C. Dolinoy (eds) *Toxicoepigenetics*. Academic Press, pp. 127–151. Available at: https://doi.org/10.1016/B978-0-12-812433-8.00005-8.

Bond, D.M. and Finnegan, E.J. (2007) 'Passing the message on: inheritance of epigenetic traits', *Trends in Plant Science*, 12(5), pp. 211–216. Available at: https://doi.org/10.1016/j.tplants.2007.03.010.

Borg, E. (2007) 'If mirror neurons are the answer, what was the question?', *Journal of Consciousness Studies*, 14(8), pp. 5–19.

Borg, E. (2013) 'More questions for mirror neurons', *Consciousness and Cognition*, 22(3), pp. 1122–1131. Available at: https://doi.org/10.1016/j.concog.2012.11.013.

Borkenau, P. and Tandler, N. (2015) 'Personality, Trait Models of', in *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, pp. 920–924. Available at: https://doi.org/10.1016/B978-0-08-097086-8.25023-X.

Bouchard, T.J. and Loehlin, J.C. (2001) 'Genes, Evolution, and Personality', *Behavior Genetics*, 31(3), pp. 243–273. Available at: https://doi.org/10.1023/A:1012294324713.

Brass, M. *et al.* (2003) 'Imitative response tendencies in patients with frontal brain lesions.', *Neuropsychology*, 17(2), pp. 265–271. Available at: https://doi.org/10.1037/0894-4105.17.2.265.

Bratko, D., Butković, A. and Hlupić, T.V. (2017) 'Heritability of Personality', *Philosophical Topics*, 26(1), p. 24.

Buchanan, R. (2012) 'Is Belief a Propositional Attitude?', *Philosopher's Imprint*, 12(1), p. 20.

Budson, A.E., Todman, R.W. and Schacter, D.L. (2006) 'Gist memory in Alzheimer's disease: Evidence from categorized pictures.', *Neuropsychology*, 20(1), pp. 113–122. Available at: https://doi.org/10.1037/0894-4105.20.1.113.

Butterfill, S.A. and Apperly, I.A. (2013) 'How to Construct a Minimal Theory of Mind', *Mind & Language*, 28(5), pp. 606–637. Available at: https://doi.org/10.1111/mila.12036.

Carey, S. (2011) *The origin of concepts*. Oxford: Oxford Univ. Press.

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

Carruthers, P. *et al.* (2012) 'The Evolution of Self-Knowledge', *Philosophical Topics*, 40(2), pp. 13–37. Available at: https://doi.org/10.5840/philtopics201240212.

Carruthers, P. (2013) 'Mindreading in Infancy', *Mind & Language*, 28(2), pp. 141–172. Available at: https://doi.org/10.1111/mila.12014.

Cattell, R.B. (1943) 'The description of personality: basic traits resolved into clusters', *Journal of Abnormal*, 38(4), pp. 476–506.

Chalmers, D.J. (2016) 'The Combination Problem for Panpsychism', in G. Bruntrup and L. Jaskolla (eds) *Panpsychism*. Oxford University Press, pp. 179–214. Available at: https://doi.org/10.1093/acprof:oso/9780199359943.003.0008.

Chartrand, T.L. and Bargh, J.A. (1999) 'The chameleon effect: The perception–behavior link and social interaction', *Journal of Personality and Social Psychology*, 76(6), pp. 893–910. Available at: https://doi.org/10.1037/0022-3514.76.6.893.

Chen, L. *et al.* (2019) 'The visual word form area (VWFA) is part of both language and attention circuitry', *Nature Communications*, 10(1), p. 5601. Available at: https://doi.org/10.1038/s41467-019-13634-z.

Choi, I., Nisbett, R.E. and Norenzayan, A. (1999) 'Causal attribution across cultures: Variation and universality.', *Psychological Bulletin*, 125(1), pp. 47–63. Available at: https://doi.org/10.1037/0033-2909.125.1.47.

Chomsky, N. (1959) 'A Review of B. F. Skinner's Verbal Behavior', *Language*, 35(1), pp. 26–58.

Christensen, W. and Michael, J. (2016) 'From two systems to a multi-systems architecture for mindreading', *New Ideas in Psychology*, 40, pp. 48–64. Available at: https://doi.org/10.1016/j.newideapsych.2015.01.003.

Churchland, P.M. (1979) *Scientific realism and the plasticity of mind*. Cambridge: Cambridge University Press.

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

Ciaramelli, E. *et al.* (2006) 'When true memory availability promotes false memory: Evidence from confabulating patients', *Neuropsychologia*, 44(10), pp. 1866–1877. Available at: https://doi.org/10.1016/j.neuropsychologia.2006.02.008.

Clark, A. (2013) 'Whatever next? Predictive brains, situated agents, and the future of cognitive science', *Behavioral and Brain Sciences*, 36(3), pp. 181–204. Available at: https://doi.org/10.1017/S0140525X12000477.

Cloninger, C.R. (1986) 'A unified biosocial theory of personality and its role in the development of anxiety states', *Psychiatric Developments*, 4(3), pp. 167–226.

Coffield, F. *et al.* (2004) 'Learning styles and pedagogy in post-16 learning: a systematic and critical review', *Learning & Skills Research Centre* [Preprint]. Available at: https://www.voced.edu.au/content/ngv%3A13692 (Accessed: 20 April 2022).

Cogsdill, E.J. *et al.* (2014) 'Inferring Character From Faces: A Developmental Study', *Psychological Science*, 25(5), pp. 1132–1139. Available at: https://doi.org/10.1177/0956797614523297.

Cohen, L. and Dehaene, S. (2004) 'Specialization within the ventral stream: the case for the visual word form area', *NeuroImage*, 22(1), pp. 466–476. Available at: https://doi.org/10.1016/j.neuroimage.2003.12.049.

Coppedè, F. (2014) 'Epigenetics and Cognitive Disorders', in *Epigenetics in Psychiatry*. Elsevier, pp. 343–367. Available at: https://doi.org/10.1016/B978-0-12-417114-5.00017-6.

Cubelli, R. (2010) 'A new taxonomy of memory and forgetting', in *Forgetting*. New York, NY, US: Psychology Press, pp. 35–47.

Curry, D.S. (2021) 'Street smarts', *Synthese*, 199(1), pp. 161–180. Available at: https://doi.org/10.1007/s11229-020-02641-z.

Darley, J.M. and Batson, C.D. (1973) '"From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior.', *Journal of Personality and Social Psychology*, 27(1), pp. 100–108. Available at: https://doi.org/10.1037/h0034449.

Day, J.J. and Sweatt, J.D. (2011) 'Epigenetic mechanisms in cognition', *Neuron*, 70(5), pp. 813–829. Available at: https://doi.org/10.1016/j.neuron.2011.05.019.

De Fruyt, F., Van De Wiele, L. and Van Heeringen, C. (2000) 'Cloninger's Psychobiological Model of Temperament and Character and the Five-Factor Model of

Personality', *Personality and Individual Differences*, 29(3), pp. 441–452. Available at: https://doi.org/10.1016/S0191-8869(99)00204-4.

Deans, C. and Maggert, K.A. (2015) 'What Do You Mean, "Epigenetic"?', *Genetics*, 199(4), pp. 887–896. Available at: https://doi.org/10.1534/genetics.114.173492.

Deese, J. (1959) 'On the prediction of occurrence of particular verbal intrusions in immediate recall.', *Journal of Experimental Psychology*, 58(1), pp. 17–22. Available at: https://doi.org/10.1037/h0046671.

Dehaene, S. and Cohen, L. (2011) 'The unique role of the visual word form area in reading', *Trends in Cognitive Sciences*, 15(6), pp. 254–262. Available at: https://doi.org/10.1016/j.tics.2011.04.003.

Delvaux, V. and Soquet, A. (2007) 'The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation', *Phonetica*, 64(2–3), pp. 145–173. Available at: https://doi.org/10.1159/000107914.

Dennett, D.C. (1978) 'Beliefs about beliefs', *Behavioral and Brain Sciences*, 1(4), pp. 568–570. Available at: https://doi.org/10.1017/S0140525X00076664.

Dennett, D.C. (1987) *The intentional stance*. Cambridge, Mass: MIT Press.

Dewhurst, J. (2017) *From Folk Psychology to Cognitive Ontology*. PhD Thesis. University of Edinburgh.

Doris, J.M. (1998) 'Persons, Situations, and Virtue Ethics', *Nous*, 32(4), pp. 504–530. Available at: https://doi.org/10.1111/0029-4624.00136.

Duverger, H. *et al.* (2007) '[Theory of mind in Asperger syndrome]', *L'Encephale*, 33(4 Pt 1), pp. 592–597.

Earp, B.D. and Trafimow, D. (2015) 'Replication, falsification, and the crisis of confidence in social psychology', *Frontiers in Psychology*, 6. Available at: https://doi.org/10.3389/fpsyg.2015.00621.

Ekman, P. *et al.* (1987) 'Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion', *Journal of personality and social psychology*, 53, pp. 712–7. Available at: https://doi.org/10.1037/0022-3514.53.4.712.

Ekman, P. (2006) *Darwin and Facial Expression: A Century of Research in Review*. ISHK.

Ekman, P. and Cordaro, D. (2011) 'What is Meant by Calling Emotions Basic', *Emotion Review*, 3(4), pp. 364–370. Available at: https://doi.org/10.1177/1754073911410740.

Ekman, P. and Friesen, W.V. (1971) 'Constants across cultures in the face and emotion', *Journal of Personality and Social Psychology*, 17(2), pp. 124–129. Available at: https://doi.org/10.1037/h0030377.

Etz, A. and Vandekerckhove, J. (2018) 'Introduction to Bayesian Inference for Psychology', *Psychonomic Bulletin & Review*, 25(1), pp. 5–34. Available at: https://doi.org/10.3758/s13423-017-1262-3.

Evans, J.S.B.T. (2010) *Thinking twice: two minds in one brain*. New York: Oxford University Press.

Evans, J.St.B.T. (2012) 'Questions and challenges for the new psychology of reasoning', *Thinking & Reasoning*, 18(1), pp. 5–31. Available at: https://doi.org/10.1080/13546783.2011.637674.

Farah, M.J., Soso, M.J. and Dasheiff, R.M. (1992) 'Visual angle of the mind's eye before and after unilateral occipital lobectomy', *Journal of Experimental Psychology. Human Perception and Performance*, 18(1), pp. 241–246. Available at: https://doi.org/10.1037//0096-1523.18.1.241.

Farroni, T. *et al.* (2005) 'Newborns' preference for face-relevant stimuli: effects of contrast polarity', *Proceedings of the National Academy of Sciences of the United States of America*, 102(47), pp. 17245–17250. Available at: https://doi.org/10.1073/pnas.0502205102.

Feinerman, O. and Korman, A. (2017) 'Individual versus collective cognition in social insects', *The Journal of Experimental Biology*, 220(Pt 1), pp. 73–82. Available at: https://doi.org/10.1242/jeb.143891.

Feldman, M.W. and Laland, K.N. (1996) 'Gene-culture coevolutionary theory', *Trends in Ecology & Evolution*, 11(11), pp. 453–457. Available at: https://doi.org/10.1016/0169-5347(96)10052-5.

Ferreira, M.B. *et al.* (2012) 'On the relation between spontaneous trait inferences and intentional inferences: An inference monitoring hypothesis', *Journal of Experimental Social Psychology*, 48(1), pp. 1–12. Available at: https://doi.org/10.1016/j.jesp.2011.06.013.

Fiebich, A. (2019) 'In defense of pluralist theory', *Synthese* [Preprint]. Available at: https://doi.org/10.1007/s11229-019-02490-5.

Fiebich, A. and Coltheart, M. (2015) 'Various Ways to Understand Other Minds: Towards a Pluralistic Approach to the Explanation of Social Understanding', *Mind and Language*, 30(3), pp. 235–258.

Fiedler, K. *et al.* (2005) 'Priming Trait Inferences Through Pictures and Moving Pictures: The Impact of Open and Closed Mindsets.', *Journal of Personality and Social Psychology*, 88(2), pp. 229–244. Available at: https://doi.org/10.1037/0022-3514.88.2.229.

Fiedler, K. and Schenck, W. (2001) 'Spontaneous Inferences from Pictorially Presented Behaviors', *Personality and Social Psychology Bulletin*, 27(11), pp. 1533–1546. Available at: https://doi.org/10.1177/01461672012711013.

Fisher, J.C. (2006) 'Does Simulation Theory Really Involve Simulation?', *Philosophical Psychology*, 19(4), pp. 417–432. Available at: https://doi.org/10.1080/09515080600726377.

Fiske, S.T. *et al.* (2002) 'A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition.', *Journal of Personality and Social Psychology*, 82(6), pp. 878–902. Available at: https://doi.org/10.1037/0022-3514.82.6.878.

Fiske, S.T., Cuddy, A.J.C. and Glick, P. (2007) 'Universal dimensions of social cognition: Warmth and competence', *Trends in Cognitive Sciences*, 11(2), pp. 77–83. Available at: https://doi.org/10.1016/j.tics.2006.11.005.

Fleeson, W. and Jayawickreme, E. (2015) 'Whole Trait Theory', *Journal of Research in Personality*, 56, pp. 82–92. Available at: https://doi.org/10.1016/j.jrp.2014.10.009.

Fodor, J.A. (1975) *The Language of Thought*. Harvard University Press.

Franklin, A., Skelton, A. and Catchpole, G. (2014) 'The case for infant colour categories', in W. Anderson et al. (eds) *Colour Studies*. Amsterdam: John Benjamins Publishing Company, pp. 169–180. Available at: https://doi.org/10.1075/z.191.11fra.

Gallagher, S. (2001) 'The Practice of Mind: Theory, Simulation or Primary Interaction?', *Journal of Consciousness Studies*, 8, pp. 83–108.

Gallagher, S. (2005) *How the Body Shapes the Mind*. Oxford University Press.

Gallagher, S. (2020) *Action and interaction*. New product. New York: Oxford University Press.

Gallese, V. *et al.* (1996) 'Action recognition in the premotor cortex', *Brain: A Journal of Neurology*, 119 (Pt 2), pp. 593–609. Available at: https://doi.org/10.1093/brain/119.2.593.

Gangopadhyay, N. and Miyahara, K. (2015) 'Perception and the problem of access to other minds', *Philosophical Psychology*, 28(5), pp. 695–714. Available at: https://doi.org/10.1080/09515089.2014.895935.

Garnier, M., Lamalle, L. and Sato, M. (2013) 'Neural correlates of phonetic convergence and speech imitation', *Frontiers in Psychology*, 4. Available at: https://doi.org/10.3389/fpsyg.2013.00600.

Gendron, M. *et al.* (2014) 'Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture.', *Emotion*, 14(2), pp. 251–262. Available at: https://doi.org/10.1037/a0036052.

Goldberg, L.R. (1993) 'The Structure of Phenotypic Personality Traits', *American Psychologist*, p. 9.

Goldman, A.I. (1989) 'Interpretation Psychologized*', *Mind & Language*, 4(3), pp. 161–185. Available at: https://doi.org/10.1111/j.1468-0017.1989.tb00249.x.

Goldman, A.I. (2006) *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Gopnik, A. and Wellman, H.M. (1994) 'The theory theory', in *Mapping the mind: Domain specificity in cognition and culture*. New York, NY, US: Cambridge University Press, pp. 257–293. Available at: https://doi.org/10.1017/CBO9780511752902.011.

Gordon, R.M. (1986) 'Folk Psychology as Simulation', *Mind & Language*, 1(2), pp. 158–171. Available at: https://doi.org/10.1111/j.1468-0017.1986.tb00324.x.

Gordon, R.M. (1995) ‘Sympathy, Simulation, and the Impartial Spectator’, *Ethics*, 105(4), pp. 727–742.

Gray, C.A. and Garand, J.D. (1993) ‘Social Stories: Improving Responses of Students with Autism with Accurate Social Information’, *Focus on Autistic Behavior*, 8(1), pp. 1–10. Available at: https://doi.org/10.1177/108835769300800101.

de Haan, M., Pascalis, O. and Johnson, M.H. (2002) ‘Specialization of neural mechanisms underlying face recognition in human infants’, *Journal of Cognitive Neuroscience*, 14(2), pp. 199–209. Available at: https://doi.org/10.1162/089892902317236849.

Hacker, P.M.S. (ed.) (2007) ‘Powers’, in *Human Nature: The Categorial Framework*. Oxford, UK: Blackwell Publishing Ltd, pp. 90–121. Available at: https://doi.org/10.1002/9780470692165.ch4.

Haith, M.M. (1998) ‘Who put the cog in infant cognition? Is rich interpretation too costly?’, *Infant Behavior and Development*, 21(2), pp. 167–179. Available at: https://doi.org/10.1016/S0163-6383(98)90001-7.

Ham, J. and Vonk, R. (2003) ‘Smart and easy: Co-occurring activation of spontaneous trait inferences and spontaneous situational inferences’, *Journal of Experimental Social Psychology*, 39(5), pp. 434–447. Available at: https://doi.org/10.1016/S0022-1031(03)00033-7.

Hampshire, S. (1953) ‘Dispositions’, *Analysis*, 14(1), pp. 5–11. Available at: https://doi.org/10.2307/3326315.

Harman, G. (1978) ‘Studying the Chimpanzee’s Theory of Mind’, *Behavioral and Brain Sciences*, 1(4), pp. 576–577. Available at: https://doi.org/10.1017/s0140525x00076743.

Harman, G. (1999) ‘XIV-Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error’, *Proceedings of the Aristotelian Society*, 99(3), pp. 315–331. Available at: https://doi.org/10.1111/1467-9264.00062.

Hassabis, D. *et al.* (2007) ‘Patients with hippocampal amnesia cannot imagine new experiences’, *Proceedings of the National Academy of Sciences*, 104(5), pp. 1726–1731. Available at: https://doi.org/10.1073/pnas.0610561104.

Hassin, R.R., Uleman, J.S. and Bargh, J.A. (2004) *The New Unconscious*. Oxford University Press.

Hayashi, H. and Nishikawa, M. (2019) ‘Egocentric bias in emotional understanding of children and adults’, *Journal of Experimental Child Psychology*, 185, pp. 224–235. Available at: https://doi.org/10.1016/j.jecp.2019.04.009.

He, Z., Bolz, M. and Baillargeon, R. (2012) ‘2.5-year-olds Succeed at a Verbal Anticipatory-Looking False-Belief Task’, *The British Journal of Developmental Psychology*, 30(Pt 1), pp. 14–29. Available at: https://doi.org/10.1111/j.2044-835X.2011.02070.x.

Heal, J. (1986) ‘Replication and Functionalism’, in J. Butterfield (ed.) *Language, Mind, and Logic*. Cambridge University Press, pp. 135–150.

Heal, J. (1996) 'Simulation and Cognitive Penetrability', *Mind & Language*, 11(1), pp. 44–67. Available at: https://doi.org/10.1111/j.1468-0017.1996.tb00028.x.

Heal, J. (1998) 'Co-Cognition and Off-Line Simulation: Two Ways of Understanding the Simulation Approach', *Mind and Language*, 13(4), pp. 477–498. Available at: https://doi.org/10.1111/1468-0017.00088.

Heberlein, A.S. *et al.* (2004) 'Cortical regions for judgments of emotions and personality traits from point-light walkers', *Journal of Cognitive Neuroscience*, 16(7), pp. 1143–1158. Available at: https://doi.org/10.1162/0898929041920423.

Henrich, J., Heine, S.J. and Norenzayan, A. (2010a) 'Most people are not WEIRD', *Nature*, 466(7302), pp. 29–29. Available at: https://doi.org/10.1038/466029a.

Henrich, J., Heine, S.J. and Norenzayan, A. (2010b) 'The weirdest people in the world?', *Behavioral and Brain Sciences*, 33(2–3), pp. 61–83. Available at: https://doi.org/10.1017/S0140525X0999152X.

Henrich, J., Heine, S.J. and Norenzayan, A. (2010c) 'The weirdest people in the world?', *Behavioral and Brain Sciences*, 33(2–3), pp. 61–83. Available at: https://doi.org/10.1017/S0140525X0999152X.

Henrich, J.P. (2020) *The WEIRDest people in the world: How the West Became Psychologically Peculiar and Particularly Prosperous*. New York: Farrar, Straus and Giroux.

Herschbach, M. (2018) *Critical Note: How Revisionary are 4E accounts of Social Cognition?*, *The Oxford Handbook of 4E Cognition*. Available at: https://doi.org/10.1093/oxfordhb/9780198735410.013.27.

Hess, U. and Blairy, S. (2001) 'Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy', *International Journal of Psychophysiology*, 40(2), pp. 129–141. Available at: https://doi.org/10.1016/S0167-8760(00)00161-6.

Heyes, C. (2012a) 'Grist and mills: on the cultural origins of cultural learning', *Philosophical Transactions of the Royal Society B: Biological Sciences* [Preprint]. Available at: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2012.0120 (Accessed: 8 July 2019).

Heyes, C. (2012b) 'New thinking: the evolution of human cognition', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599), pp. 2091–2096. Available at: https://doi.org/10.1098/rstb.2012.0111.

Heyes, C. (2014) 'False belief in infancy: a fresh look', *Developmental Science*, 17(5), pp. 647–659. Available at: https://doi.org/10.1111/desc.12148.

Heyes, C. (2019a) 'Précis of Cognitive Gadgets: The Cultural Evolution of Thinking', *Behavioral and Brain Sciences*, 42. Available at: https://doi.org/10.1017/S0140525X18002145.

Heyes, C. (2019b) *Testing cognitive gadgets*, *Mind & Language*. Available at: https://doi.org/10.1111/mila.12253.

Heyes, C.M. (2018) *Cognitive gadgets: the cultural evolution of thinking*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Hurley, S. (2008) 'Understanding Simulation', *Philosophy and Phenomenological Research*, 77(3), pp. 755–774. Available at: https://doi.org/10.1111/j.1933-1592.2008.00220.x.

Hutto, D.D. (2004) 'The Limits of Spectatorial Folk Psychology', *Mind & Language*, 19(5), pp. 548–573. Available at: https://doi.org/10.1111/j.0268-1064.2004.00272.x.

Hutto, D.D. (2007) 'The Narrative Practice Hypothesis: Origins and Applications of Folk Psychology', *Royal Institute of Philosophy Supplements*, 60, pp. 43–68. Available at: https://doi.org/10.1017/S1358246107000033.

Hutto, D.D. (2008a) *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Bradford.

Hutto, D.D. (2008b) 'The Narrative Practice Hypothesis: Clarifications and Implications', *Philosophical Explorations*, 11(3), pp. 175–192.

Ickes, W. (2011) 'Everyday Mind Reading Is Driven by Motives and Goals', *Psychological Inquiry*, 22(3), pp. 200–206. Available at: https://doi.org/10.1080/1047840X.2011.561133.

Izard, C.E. (1971) *The face of emotion*. East Norwalk, CT, US: Appleton-Century-Crofts (The face of emotion), pp. xii, 468.

Jack, R.E. *et al.* (2009) 'Cultural confusions show that facial expressions are not universal', *Current biology*, 19(18), pp. 1543–1548.

Jack, R.E. *et al.* (2012) 'Facial expressions of emotion are not culturally universal', *Proceedings of the National Academy of Sciences*, 109(19), pp. 7241–7244. Available at: https://doi.org/10.1073/pnas.1200155109.

Jackson, F. (1999) 'All that can be at issue in the theory-theory simulation debate', *Philosophical Papers*, 28(2), pp. 77–96. Available at: https://doi.org/10.1080/05568649909506593.

John, L.K., Loewenstein, G. and Prelec, D. (2012) 'Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling', *Psychological Science*, 23(5), pp. 524–532. Available at: https://doi.org/10.1177/0956797611430953.

Johnson, M.H. *et al.* (1991) 'Newborns' preferential tracking of face-like stimuli and its subsequent decline', *Cognition*, 40(1–2), pp. 1–19. Available at: https://doi.org/10.1016/0010-0277(91)90045-6.

Johnson, M.H. (2005) 'Subcortical face processing', *Nature Reviews Neuroscience*, 6(10), pp. 766–774. Available at: https://doi.org/10.1038/nrn1766.

Johnson, S.C. *et al.* (2002) 'Neural correlates of self-reflection', *Brain*, 125(8), pp. 1808–1814. Available at: https://doi.org/10.1093/brain/awf181.

Jones, E.E. and Harris, V.A. (1967) 'The attribution of attitudes', *Journal of Experimental Social Psychology*, 3(1), pp. 1–24. Available at: https://doi.org/10.1016/0022-1031(67)90034-0.

Kahneman, D. (2011) *Thinking, fast and slow*. 1st ed. New York: Farrar, Straus and Giroux.

Kalish, C.W. (2002) 'Children's predictions of consistency in people's actions', *Cognition*, 84(3), pp. 237–265. Available at: https://doi.org/10.1016/S0010-0277(02)00052-5.

Kappes, J. *et al.* (2009) 'Unintended imitation in nonword repetition', *Brain and Language*, 111(3), pp. 140–151. Available at: https://doi.org/10.1016/j.bandl.2009.08.008.

Kay, P. and Kempton, W. (1984) 'What Is the Sapir-Whorf Hypothesis?', *American Anthropologist*, 86(1), pp. 65–79. Available at: https://doi.org/10.1525/aa.1984.86.1.02a00050.

Kayyal, M., Widen, S. and Russell, J.A. (2015) 'Context is more powerful than we think: Contextual cues override facial cues even for valence.', *Emotion*, 15(3), pp. 287–291. Available at: https://doi.org/10.1037/emo0000032.

Kazak, S., Collis, G.M. and Lewis, V. (1997) 'Can Young People with Autism Refer to Knowledge States? Evidence from Their Understanding of "Know" and "Guess"', *Journal of Child Psychology and Psychiatry*, 38(8), pp. 1001–1009. Available at: https://doi.org/10.1111/j.1469-7610.1997.tb01617.x.

Kelley, W.M. *et al.* (2002) 'Finding the self? An event-related fMRI study', *Journal of Cognitive Neuroscience*, 14(5), pp. 785–794. Available at: https://doi.org/10.1162/08989290260138672.

Klein, S.B., Loftus, J. and Kihlstrom, J.F. (2002) 'Memory and Temporal Experience: the Effects of Episodic Memory Loss on an Amnesic Patient's Ability to Remember the Past and Imagine the Future', *Social Cognition*, 20(5), pp. 353–379. Available at: https://doi.org/10.1521/soco.20.5.353.21125.

Knoblich, G. and Sebanz, N. (2016) 'The Social Nature of Perception and Action':, *Current Directions in Psychological Science* [Preprint]. Available at: http://journals.sagepub.com/doi/10.1111/j.0963-7214.2006.00415.x (Accessed: 16 November 2020).

Knudsen, B. and Liszkowski, U. (2012) 'Eighteen- and 24-month-old infants correct others in anticipation of action mistakes', *Developmental Science*, 15(1), pp. 113–122. Available at: https://doi.org/10.1111/j.1467-7687.2011.01098.x.

Kokina, A. and Kern, L. (2010) 'Social Story^{TM} Interventions for Students with Autism Spectrum Disorders: A Meta-Analysis', *Journal of Autism and Developmental Disorders*, 40(7), pp. 812–826. Available at: https://doi.org/10.1007/s10803-009-0931-0.

Kutas, M. and Federmeier, K.D. (2011) 'Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP)', *Annual review of psychology*, 62, pp. 621–647. Available at: https://doi.org/10.1146/annurev.psych.093008.131123.

Kutas, M. and Hillyard, S.A. (1980) 'Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity', *Science*, 207(4427), pp. 203–205.

Kutas, M. and Hillyard, S.A. (1984) 'Brain potentials during reading reflect word expectancy and semantic association', *Nature*, 307(5947), pp. 161–163. Available at: https://doi.org/10.1038/307161a0.

Kuwabara, M., Son, J.Y. and Smith, L.B. (2011) 'Attention to Context: U.S. and Japanese Children's Emotional Judgments', *Journal of Cognition and Development*, 12(4), pp. 502–517. Available at: https://doi.org/10.1080/15248372.2011.554927.

Kwong, J.M.C. (2007) 'Is Conceptual Atomism a Plausible Theory of Concepts?', *The Southern Journal of Philosophy*, 45(3), pp. 413–434. Available at: https://doi.org/10.1111/j.2041-6962.2007.tb00058.x.

Lakin, J.L. *et al.* (2003) 'The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry', *Journal of Nonverbal Behavior*, 27(3), pp. 145–162. Available at: https://doi.org/10.1023/A:1025389814290.

Lalande, M. and Calciano, M.A. (2007) 'Molecular epigenetics of Angelman syndrome', *Cellular and molecular life sciences: CMLS*, 64(7–8), pp. 947–960. Available at: https://doi.org/10.1007/s00018-007-6460-0.

Lamiell, J.T. (2018) 'On the concepts of character and personality: Correctly interpreting the statistical evidence putatively relevant to the disposition–situation debate', *Theory & Psychology*, 28(2), pp. 249–254. Available at: https://doi.org/10.1177/0959354317748374.

Laurence, S. and Margolis, E. (2001) 'The Poverty of the Stimulus Argument', *British Journal for the Philosophy of Science*, 52(2), pp. 217–276. Available at: https://doi.org/10.1093/bjps/52.2.217.

Lavelle, J.S. (2012) 'Two Challenges to Hutto's Enactive Account of Pre-linguistic Social Cognition', *Philosophia*, 40(3), pp. 459–472. Available at: https://doi.org/10.1007/s11406-011-9356-z.

Lavelle, J.S. (2019) *The social mind: a philosophical introduction*. London New York: Routledge.

Lavelle, J.S. (2021) 'The impact of culture on mindreading', *Synthese*, 198(7), pp. 6351–6374. Available at: https://doi.org/10.1007/s11229-019-02466-5.

Leighton, J. *et al.* (2010) 'Social attitudes modulate automatic imitation', *Journal of Experimental Social Psychology*, 46(6), pp. 905–910. Available at: https://doi.org/10.1016/j.jesp.2010.07.001.

Levordashka, A. and Utz, S. (2017) 'Spontaneous Trait Inferences on Social Media', *Social Psychological and Personality Science*, 8(1), pp. 93–101. Available at: https://doi.org/10.1177/1948550616663803.

Lewis, D. (1972) 'Psychophysical and theoretical identifications', *Australasian Journal of Philosophy*, 50(3), pp. 249–258.

Loftus, E.F. (1975) 'Leading questions and the eyewitness report', *Cognitive Psychology*, 7(4), pp. 560–572. Available at: https://doi.org/10.1016/0010-0285(75)90023-7.

Loftus, E.F. (2005) 'Planting misinformation in the human mind: a 30-year investigation of the malleability of memory', *Learning & Memory (Cold Spring Harbor, N.Y.)*, 12(4), pp. 361–366. Available at: https://doi.org/10.1101/lm.94705.

Lucas, R.E. and Donnellan, M.B. (2011) 'Personality development across the life span: Longitudinal analyses with a national sample from Germany', *Journal of Personality and Social Psychology*, 101(4), pp. 847–861. Available at: https://doi.org/10.1037/a0024298.

Lucca, K., Hamlin, J.K. and Sommerville, J. (eds) (2019) *Early Moral Cognition and Behavior*. Frontiers Media SA (Frontiers Research Topics). Available at: https://doi.org/10.3389/978-2-88963-188-9.

Lux, V. (2013) 'With Gottlieb beyond Gottlieb: The Role of Epigenetics in Psychobiological Development', in. Available at: https://doi.org/10.3233/DEV-1300073.

Lycan, W.G. (2012) 'Desire considered as a propositional attitude', *Philosophical Perspectives*, 26, pp. 201–215.

Maibom, H. (2017) *The Routledge Handbook of Philosophy of Empathy*. Taylor & Francis.

Malle, B.F. (1999) 'How People Explain Behavior: A New Theoretical Framework', *Personality and Social Psychology Review*, 3(1), pp. 23–48. Available at: https://doi.org/10.1207/s15327957pspr0301_2.

Malle, B.F. (2004) *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA, US: MIT Press (How the mind explains behavior: Folk explanations, meaning, and social interaction).

Malle, B.F. (2011) 'Time to Give Up the Dogmas of Attribution', in *Advances in Experimental Social Psychology*. Elsevier, pp. 297–352. Available at: https://doi.org/10.1016/B978-0-12-385522-0.00006-8.

Margolis, E. and Laurence, S. (eds) (1999) *Concepts: core readings*. Cambridge, Mass: MIT Press.

Marshall, P. and Bredy, T.W. (2016) 'Cognitive neuroepigenetics: the next evolution in our understanding of the molecular mechanisms underlying learning and memory?', *npj Science of Learning*, 1(1), pp. 1–8. Available at: https://doi.org/10.1038/npjscilearn.2016.14.

Matsumoto, D. and Ekman, P. (1989) 'American-Japanese cultural differences in intensity ratings of facial expressions of emotion', *Motivation and Emotion*, 13(2), pp. 143–157. Available at: https://doi.org/10.1007/BF00992959.

Matthen, M. (2014) 'Debunking enactivism: a critical notice of Hutto and Myin's Radicalizing Enactivism', *Canadian Journal of Philosophy*, 44(1), pp. 118–128. Available at: https://doi.org/10.1080/00455091.2014.905251.

Matthews, G. (2015) 'Personality, Cognitive Models of', in *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, pp. 870–875. Available at: https://doi.org/10.1016/B978-0-08-097086-8.25075-7.

Matthews, G. (2018) 'Cognitive-Adaptive Trait Theory: A Shift in Perspective on Personality', *Journal of Personality*, 86(1), pp. 69–82. Available at: https://doi.org/10.1111/jopy.12319.

McCrae, R.R. *et al.* (1998) 'Cross-Cultural Assessment of the Five-Factor Model: The Revised NEO Personality Inventory', *Journal of Cross-Cultural Psychology*, 29(1), pp. 171–188. Available at: https://doi.org/10.1177/0022022198291009.

McCrae, R.R. and Costa, P.T. (1987) 'Validation of the Five-Factor Model of Personality Across Instruments and Observers', pp. 81–90.

McCrae, R.R. and Costa, P.T. (2003) *Personality in Adulthood: A Five-factor Theory Perspective*. Guilford Press.

Melis, A.P., Hare, B. and Tomasello, M. (2006) 'Chimpanzees Recruit the Best Collaborators', *Science*, 311(5765), pp. 1297–1300. Available at: https://doi.org/10.1126/science.1123007.

Mellor, D.H. (2000) 'The Semantics and Ontology of Dispositions', *Mind*, 109(436), pp. 757–780.

Mendiola, A.J.P. and LaSalle, J.M. (2021) 'Epigenetics in Prader-Willi Syndrome', *Frontiers in Genetics*, 12. Available at: https://www.frontiersin.org/article/10.3389/fgene.2021.624581 (Accessed: 26 May 2022).

Merricks, T. (2009) 'Propositional Attitudes?', *Proceedings of the Aristotelian Society*, 109, pp. 207–232.

Michael, J. (2011) 'Interactionism and Mindreading', *Review of Philosophy and Psychology*, 2(3), pp. 559–578. Available at: https://doi.org/10.1007/s13164-011-0066-z.

Milkowski, M. (2013) *Explaining the Computational Mind*. MIT Press.

Miłkowski, M. (2015) 'The Hard Problem Of Content: Solved (Long Ago)', *Studies in Logic, Grammar and Rhetoric*, 41(1), pp. 73–88. Available at: https://doi.org/10.1515/slgr-2015-0021.

Miller, C.B. (2018) 'Generosity: A Preliminary Account of a Surprisingly Neglected Virtue', *Metaphilosophy*, 49(3), pp. 216–245. Available at: https://doi.org/10.1111/meta.12298.

Miller, J.G. (1984) 'Culture and the development of everyday social explanation.', *Journal of Personality and Social Psychology*, 46(5), pp. 961–978. Available at: https://doi.org/10.1037/0022-3514.46.5.961.

Millidge, B., Seth, A. and Buckley, C.L. (2021) 'Predictive Coding: a Theoretical and Experimental Review', *arXiv:2107.12979 [cs, q-bio]* [Preprint]. Available at: http://arxiv.org/abs/2107.12979 (Accessed: 3 January 2022).

Miozzo, M. and Caramazza, A. (1998) 'Varieties of Pure Alexia: The Case of Failure to Access Graphemic Representations', *Cognitive Neuropsychology*, 15(1–2), pp. 203–238. Available at: https://doi.org/10.1080/026432998381267.

Mitchell, J.P., Banaji, M.R. and Macrae, C.N. (2005) 'The link between social cognition and self-referential thought in the medial prefrontal cortex', *Journal of Cognitive Neuroscience*, 17(8), pp. 1306–1315. Available at: https://doi.org/10.1162/0898929055002418.

Mitchell, J.P., Macrae, C.N. and Banaji, M.R. (2004) 'Encoding-Specific Effects of Social Cognition on the Neural Correlates of Subsequent Memory', *The Journal of Neuroscience*, 24(21), pp. 4912–4917. Available at: https://doi.org/10.1523/JNEUROSCI.0481-04.2004.

Monk, M. (1995) 'Epigenetic programming of differential gene expression in development and evolution', *Developmental Genetics*, 17(3), pp. 188–197. Available at: https://doi.org/10.1002/dvg.1020170303.

Morewedge, C.K., Gilbert, D.T. and Wilson, T.D. (2005) 'The Least Likely of Times: How Remembering the Past Biases Forecasts of the Future', *Psychological Science*, 16(8), pp. 626–630. Available at: https://doi.org/10.1111/j.1467-9280.2005.01585.x.

Morin, O. (2019) 'Did social cognition evolve by cultural group selection?', *Mind & Language*, 0(0). Available at: https://doi.org/10.1111/mila.12252.

Mumford, S. (2003) *Dispositions*. Oxford, New York: Oxford University Press.

Mumford, S. and Anjum, R.L. (2011) *Getting causes from powers*. Oxford ; New York: Oxford University Press.

Munakata, Y. (2000) 'Challenges to the Violation-of-Expectation Paradigm: Throwing the Conceptual Baby Out With the Perceptual Processing Bathwater?', *Infancy*, 1(4), pp. 471–477. Available at: https://doi.org/10.1207/S15327078IN0104_7.

Myers, I.B. and Briggs, K.C. (1962) *The Myers-Briggs type indicator : manual (1962)*. Palo Alto, Calif. : Consulting Psychologists Press [distributor]. Available at: https://trove.nla.gov.au/version/250954134 (Accessed: 1 February 2019).

Na, J. and Kitayama, S. (2011) 'Spontaneous Trait Inference Is Culture-Specific: Behavioral and Neural Evidence', *Psychological Science*, 22(8), pp. 1025–1032. Available at: https://doi.org/10.1177/0956797611414727.

Nadelhoffer, T., Nahmias, E. and Nichols, S. (2010) *Moral Psychology: Historical and Contemporary Readings*. Wiley.

Nagel, T. (1979) 'Panpsychism', in T. Nagel (ed.) *Mortal Questions*. Cambridge University Press.

Nelson, N.L. and Russell, J.A. (2013) 'Universality Revisited', *Emotion Review*, 5(1), pp. 8–15. Available at: https://doi.org/10.1177/1754073912457227.

Neumann, O. (1984) 'Automatic Processing: A Review of Recent Findings and a Plea for an Old Theory', in W. Prinz and A.F. Sanders (eds) *Cognition and Motor Processes*.

Berlin, Heidelberg: Springer, pp. 255–293. Available at: https://doi.org/10.1007/978-3-642-69382-3_17.

Newen, A., Bruin, L. de and Gallagher, S. (eds) (2020) *The Oxford handbook of 4E cognition*. First published in paperback. Oxford: Oxford University Press.

Newman, L.S. (1991) 'Why Are Traits Inferred Spontaneously? A Developmental Approach', *Social Cognition*, 9(3), pp. 221–253. Available at: https://doi.org/10.1521/soco.1991.9.3.221.

Nichols, S. and Stich, S.P. (2003) *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford University Press.

Nichols, S.L. *et al.* (2015) 'Review of social story interventions for children diagnosed with autism spectrum disorders.', in *Journal of Evidence Based Practice for Schools*. Rowman & Littlefield (1).

O'Connell, G. *et al.* (2018) 'Thinking about others and the future: Neural correlates of perspective taking relate to preferences for delayed rewards', *Cognitive, Affective, & Behavioral Neuroscience*, 18(1), pp. 35–42. Available at: https://doi.org/10.3758/s13415-017-0550-8.

O'Connell, G., Christakou, A. and Chakrabarti, B. (2015) 'The role of simulation in intertemporal choices', *Frontiers in Neuroscience*, 9, p. 94. Available at: https://doi.org/10.3389/fnins.2015.00094.

Okuda, J. *et al.* (2003) 'Thinking of the future and past: the roles of the frontal pole and the medial temporal lobes', *NeuroImage*, 19(4), pp. 1369–1380. Available at: https://doi.org/10.1016/S1053-8119(03)00179-4.

Onishi, K.H. and Baillargeon, R. (2005) 'Do 15-Month-Old Infants Understand False Beliefs?', *Science (New York, N.y.)*, 308(5719), pp. 255–258. Available at: https://doi.org/10.1126/science.1107621.

Open Science Collaboration (2015) 'Estimating the reproducibility of psychological science', *Science*, 349(6251), p. aac4716. Available at: https://doi.org/10.1126/science.aac4716.

Over, H. and Cook, R. (2018) 'Where do spontaneous first impressions of faces come from?', *Cognition*, 170, pp. 190–200. Available at: https://doi.org/10.1016/j.cognition.2017.10.002.

Pascalis, O. *et al.* (1995) 'Mother's face recognition by neonates: A replication and an extension', *Infant Behavior and Development*, 18(1), pp. 79–85. Available at: https://doi.org/10.1016/0163-6383(95)90009-8.

Passino, K.M., Seeley, T.D. and Visscher, P.K. (2008) 'Swarm cognition in honey bees', *Behavioral Ecology and Sociobiology*, 62(3), pp. 401–414. Available at: https://doi.org/10.1007/s00265-007-0468-1.

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

Peacocke, C. (2008) *Truly understood*. Oxford ; New York: Oxford University Press.

Peteron, E.J. (2021) *What can astrocytes compute?*, p. 2021.10.20.465192. Available at: https://doi.org/10.1101/2021.10.20.465192.

Piccinini, G. (2015) *Physical computation: a mechanistic account*. Oxford: Oxford University Press.

Pickering, M.J. and Garrod, S. (2007) 'Do people use language production to make predictions during comprehension?', *Trends in Cognitive Sciences*, 11(3), pp. 105–110. Available at: https://doi.org/10.1016/j.tics.2006.12.002.

Ponsi, G. *et al.* (2016) 'Influence of warmth and competence on the promotion of safe in-group selection: Stereotype content model and social categorization of faces', *The Quarterly Journal of Experimental Psychology*, 69(8), pp. 1464–1479. Available at: https://doi.org/10.1080/17470218.2015.1084339.

Power, R.A. and Pluess, M. (2015) 'Heritability estimates of the Big Five personality traits based on common genetic variants', *Translational Psychiatry*, 5(7), pp. 1–4. Available at: https://doi.org/10.1038/tp.2015.96.

Premack, D. and Woodruff, G. (1978) 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences*, 1(04), p. 515. Available at: https://doi.org/10.1017/S0140525X00076512.

Price, C.J. and Devlin, J.T. (2003) 'The myth of the visual word form area', *NeuroImage*, 19(3), pp. 473–481. Available at: https://doi.org/10.1016/S1053-8119(03)00084-3.

Pylyshyn, Z.W. (1978) 'When is attribution of beliefs justified? [P&W]', *Behavioral and Brain Sciences*, 1(4), pp. 592–593. Available at: https://doi.org/10.1017/S0140525X00076895.

Pylyshyn, Z.W. (1980) 'Computation and cognition: issues in the foundations of cognitive science', *Behavioral and Brain Sciences*, 3(01), p. 111. Available at: https://doi.org/10.1017/S0140525X00002053.

Ratcliffe, M. (2007) 'Commonsense Psychology, Theory of Mind and Simulation', in M. Ratcliffe (ed.) *Rethinking Commonsense Psychology: A Critique of Folk Psychology, Theory of Mind and Simulation*. London: Palgrave Macmillan UK (New Directions in Philosophy and Cognitive Science), pp. 1–26. Available at: https://doi.org/10.1007/978-0-230-62529-7_1.

Ravenscroft, I. (2016) *Folk Psychology as a Theory (Stanford Encyclopedia of Philosophy/Winter 2016 Edition)*. Available at: https://stanford.library.sydney.edu.au/archives/win2016/entries/folkpsych-theory/ (Accessed: 20 March 2020).

Ray, E. and Heyes, C. (2011) 'Imitation in infancy: the wealth of the stimulus', *Developmental Science*, 14(1), pp. 92–105. Available at: https://doi.org/10.1111/j.1467-7687.2010.00961.x.

Reid, V.M. *et al.* (2017) 'The Human Fetus Preferentially Engages with Face-like Visual Stimuli', *Current Biology*, 27(12), pp. 1825-1828.e3. Available at: https://doi.org/10.1016/j.cub.2017.05.044.

Renner, M. (1960) 'The contribution of the honey bee to the study of time-sense and astronomical orientation', *Cold Spring Harbor Symposia on Quantitative Biology*, 25, pp. 361–367. Available at: https://doi.org/10.1101/sqb.1960.025.01.037.

Repacholi, B.M. *et al.* (2016) 'Infants' generalizations about other people's emotions: Foundations for trait-like attributions.', *Developmental Psychology*, 52(3), pp. 364–378. Available at: https://doi.org/10.1037/dev0000097.

Rholes, W., Newman, L.S. and Ruble, D.N. (1990) 'Understanding self and other: Developmental and motivational aspects of perceiving persons in terms of invariant dispositions.', in *Handbook of motivation and cognition: Foundations of social behavior*. New York: Guilford Press, pp. 369–407.

Rim, S., Uleman, J.S. and Trope, Y. (2009) 'Spontaneous trait inference and construal level theory: Psychological distance increases nonconscious trait thinking', *Journal of Experimental Social Psychology*, 45(5), pp. 1088–1097. Available at: https://doi.org/10.1016/j.jesp.2009.06.015.

Rizzolatti, G. *et al.* (1996) 'Premotor cortex and the recognition of motor actions', *Cognitive Brain Research*, pp. 131–141.

Rizzolatti, G., Fogassi, L. and Gallese, V. (2001) 'Neurophysiological mechanisms underlying the understanding and imitation of action', *Nature Reviews Neuroscience*, 2(9), pp. 661–670. Available at: https://doi.org/10.1038/35090060.

Rizzolatti, G. and Sinigaglia, C. (2016) 'The mirror mechanism: a basic principle of brain function', *Nature Reviews. Neuroscience*, 17(12), pp. 757–765. Available at: https://doi.org/10.1038/nrn.2016.135.

Roberson, D. and Davidoff, J. (2000) 'The categorical perception of colors and facial expressions: The effect of verbal interference', *Memory & Cognition*, 28(6), pp. 977–986. Available at: https://doi.org/10.3758/BF03209345.

Roediger, H.L. and McDermott, K.B. (1995) 'Creating false memories: Remembering words not presented in lists.', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4), pp. 803–814. Available at: https://doi.org/10.1037/0278-7393.21.4.803.

Roige, A. and Carruthers, P. (2019) 'Cognitive instincts versus cognitive gadgets: A fallacy', *Mind & Language*, p. mila.12251. Available at: https://doi.org/10.1111/mila.12251.

Rosenbaum, R.S. *et al.* (2005) 'The case of K.C.: contributions of a memory-impaired person to memory theory', *Neuropsychologia*, 43(7), pp. 989–1021. Available at: https://doi.org/10.1016/j.neuropsychologia.2004.10.007.

Russell, J.A. (2016) 'A Sceptical Look at Faces as Emotion Signals', in C. Abell and J. Smith (eds) *The Expression of Emotion*. Cambridge: Cambridge University Press, pp. 157–172. Available at: https://doi.org/10.1017/CBO9781316275672.008.

Sadeh, N. and Verona, E. (2008) 'Psychopathic Personality Traits Associated with Abnormal Selective Attention and Impaired Cognitive Control', *Neuropsychology*, 22(5), pp. 669–680. Available at: https://doi.org/10.1037/a0012692.

Sansosti, F.J., Powell-Smith, K.A. and Kincaid, D. (2004) 'A Research Synthesis of Social Story Interventions for Children With Autism Spectrum Disorders', *Focus on Autism and Other Developmental Disabilities*, 19(4), pp. 194–204. Available at: https://doi.org/10.1177/10883576040190040101.

Sartre, J.-P. (1996) *Being and nothingness*. Translated by H.E. Barnes. London: Routledge.

Schacter, D.L. and Addis, D.R. (2007) 'The cognitive neuroscience of constructive memory: remembering the past and imagining the future', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481), pp. 773–786. Available at: https://doi.org/10.1098/rstb.2007.2087.

Schacter, D.L. and Addis, D.R. (2009) 'Remembering the Past to Imagine the Future: A Cognitive Neuroscience Perspective', *Military Psychology*, 21(sup1), pp. S108–S112. Available at: https://doi.org/10.1080/08995600802554748.

Schmitz, T.W., Kawahara-Baccus, T.N. and Johnson, S.C. (2004) 'Metacognitive evaluation, self-relevance, and the right prefrontal cortex', *NeuroImage*, 22(2), pp. 941–947. Available at: https://doi.org/10.1016/j.neuroimage.2004.02.018.

Scott, R.M. and Baillargeon, R. (2009) 'Which Penguin Is This? Attributing False Beliefs About Object Identity at 18 Months', *Child Development*, 80(4), pp. 1172–1196. Available at: https://doi.org/10.1111/j.1467-8624.2009.01324.x.

Segall, M.H., Campbell, D.T. and Herskovits, M.J. (1966) *The influence of culture on visual perception*. Indianapolis: Bobbs-Merrill Co.

Senju, A. *et al.* (2011) 'Do 18-months-olds really attribute mental states to others? A critical test', *Psychological science*, 22(7), pp. 878–880. Available at: https://doi.org/10.1177/0956797611411584.

Senju, A. (2012) 'Spontaneous theory of mind and its absence in autism spectrum disorders', *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 18(2), pp. 108–113. Available at: https://doi.org/10.1177/1073858410397208.

Shea, N. (2013) 'Naturalising Representational Content', *Philosophy Compass*, 8(5), pp. 496–509. Available at: https://doi.org/10.1111/phc3.12033.

Shockley, K., Santana, M.-V. and Fowler, C.A. (2003) 'Mutual interpersonal postural constraints are involved in cooperative conversation.', *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), pp. 326–332. Available at: https://doi.org/10.1037/0096-1523.29.2.326.

Southgate, V. and Vernetti, A. (2014) 'Belief-based action prediction in preverbal infants', *Cognition*, 130(1), pp. 1–10. Available at: https://doi.org/10.1016/j.cognition.2013.08.008.

Spaulding, S. (2010a) 'Embodied Cognition and Mindreading', *Mind & Language*, 25(1), pp. 119–140. Available at: https://doi.org/10.1111/j.1468-0017.2009.01383.x.

Spaulding, S. (2010b) ‘Embodied Cognition and Theory of Mind’, in *The Routledge Handbook of Embodied Cognition*. Routledge. Available at: https://doi.org/10.4324/9781315775845.ch19.

Spaulding, S. (2015) ‘Phenomenology of Social Cognition’, *Erkenntnis*, 80(5), pp. 1069–1089. Available at: https://doi.org/10.1007/s10670-014-9698-6.

Spaulding, S. (2016) ‘Mind Misreading’, *Philosophical Issues*, 26(1), pp. 422–440. Available at: https://doi.org/10.1111/phis.12070.

Spaulding, S. (2018a) *How We Understand Others: Philosophy and Social Cognition*. Milton, UK: Routledge.

Spaulding, S. (2018b) ‘Mindreading beyond belief: A more comprehensive conception of how we understand others’, *Philosophy Compass*, 13(11), p. 11. Available at: https://doi.org/10.1111/phc3.12526.

Spengler, S., von Cramon, D.Y. and Brass, M. (2010) ‘Resisting motor mimicry: Control of imitation involves processes central to social cognition in patients with frontal and temporo-parietal lesions’, *Social Neuroscience*, 5(4), pp. 401–416. Available at: https://doi.org/10.1080/17470911003687905.

Spivey, M.J. *et al.* (2000) ‘Eye movements during comprehension of spoken scene descriptions’, in *Proceedings of the 22nd Annual Conference of the Cognitive Science Society (pp. 487–492). Mahwah, NJ: Lawrence Erlbaum Associates Inc.*

Stagner, R. (1977) ‘On the reality and relevance of traits’, *Journal of General Psychology; Provincetown, Mass., etc.*, 96, pp. 185–207.

Stanley, J. and Williamson, T. (2017) ‘Skill’, *Noûs*, 51(4), pp. 713–726. Available at: https://doi.org/10.1111/nous.12144.

Stich, S. and Nichols, S. (1997) ‘Cognitive Penetrability, Rationality and Restricted Simulation’, *Mind & Language*, 12(3–4), pp. 297–326. Available at: https://doi.org/10.1111/j.1468-0017.1997.tb00076.x.

Stich, S.P. (1998) *Deconstructing the Mind*. Oxford University Press.

Stokes, D. (2013) ‘Cognitive Penetrability of Perception’, *Philosophy Compass*, 8(7), pp. 646–663. Available at: https://doi.org/10.1111/phc3.12043.

Stroop, J.R. (1935) ‘Studies of interference in serial verbal reactions.’, *Journal of Experimental Psychology*, 18(6), pp. 643–662. Available at: https://doi.org/10.1037/h0054651.

Suddendorf, T. and Corballis, M.C. (1997) ‘Mental time travel and the evolution of the human mind’, *Genetic, Social, and General Psychology Monographs*, 123(2), pp. 133–167.

Szpunar, K.K., Watson, J.M. and McDermott, K.B. (2007) ‘Neural substrates of envisioning the future’, *Proceedings of the National Academy of Sciences*, 104(2), pp. 642–647. Available at: https://doi.org/10.1073/pnas.0610082104.

Tajfel, H. and Turner, J.C. (2010) *An integrative theory of intergroup conflict*. New York, NY, US: Psychology Press (Rediscovering social identity).

Test, D.W. *et al.* (2011) 'A Comprehensive Review and Meta-Analysis of the Social Stories Literature', *Focus on Autism and Other Developmental Disabilities*, 26(1), pp. 49–62. Available at: https://doi.org/10.1177/1088357609351573.

Thierry, G. *et al.* (2009) 'Unconscious effects of language-specific terminology on preattentive color perception', *Proceedings of the National Academy of Sciences of the United States of America*, 106(11), pp. 4567–4570. Available at: https://doi.org/10.1073/pnas.0811155106.

Thompson, J.R. (2014) 'Signature Limits in Mindreading Systems', *Cognitive Science*, 38(7), pp. 1432–1455. Available at: https://doi.org/10.1111/cogs.12117.

Thomson, J.J. (2008) *Normativity*. Chicago, IL, UNITED STATES: Open Court. Available at: http://ebookcentral.proquest.com/lib/manchester/detail.action?docID=4206329 (Accessed: 30 March 2022).

Todorov, A. *et al.* (2008) 'Understanding evaluation of faces on social dimensions', *Trends in Cognitive Sciences*, 12(12), pp. 455–460. Available at: https://doi.org/10.1016/j.tics.2008.10.001.

Tomonaga, M. *et al.* (2004) 'Development of social cognition in infant chimpanzees (Pan troglodytes): Face recognition, smiling, gaze, and the lack of triadic interactions: Development of social cognition in infant chimpanzees', *Japanese Psychological Research*, 46(3), pp. 227–235. Available at: https://doi.org/10.1111/j.1468-5584.2004.00254.x.

Träuble, B., Marinović, V. and Pauen, S. (2010) 'Early Theory of Mind Competencies: Do Infants Understand Others' Beliefs?', *Infancy*, 15(4), pp. 434–444. Available at: https://doi.org/10.1111/j.1532-7078.2009.00025.x.

Tulving, E. (1972) 'Episodic and semantic memory', in *Organization of memory*. Oxford, England: Academic Press, pp. xiii, 423–xiii, 423.

Tulving, E. (1985) 'Memory and consciousness', *Canadian Psychology/Psychologie canadienne*, 26(1), pp. 1–12. Available at: https://doi.org/10.1037/h0080017.

Tulving, E. *et al.* (1988) 'Priming of semantic autobiographical knowledge: A case study of retrograde amnesia', *Brain and Cognition*, 8(1), pp. 3–20. Available at: https://doi.org/10.1016/0278-2626(88)90035-8.

Tulving, E. and Thomson, D.M. (1973) 'Encoding Specificity and Retrieval Processes in Episodic Memory', *Psychological Review*, 80(5), pp. 352–373.

Uleman, J.S., Adil Saribay, S. and Gonzalez, C.M. (2008) 'Spontaneous Inferences, Implicit Impressions, and Implicit Theories', *Annual Review of Psychology*, 59(1), pp. 329–360. Available at: https://doi.org/10.1146/annurev.psych.59.103006.093707.

Uleman, J.S. and Moskowitz, G.B. (1994) 'Unintended effects of goals on unintended inferences', *Journal of Personality and Social Psychology*, 66(3), pp. 490–501. Available at: https://doi.org/10.1037//0022-3514.66.3.490.

Uleman, J.S., Newman, L. and Winter, L. (1992) 'Can personality traits be inferred automatically? Spontaneous inferences require cognitive capacity at encoding', *Consciousness and Cognition*, 1(1), pp. 77–90. Available at: https://doi.org/10.1016/1053-8100(92)90049-G.

Valentine, T. and Ferrara, A. (1991) 'Typicality in categorization, recognition and identification: Evidence from face recognition', *British Journal of Psychology*, 82(1), p. 87. Available at: https://doi.org/10.1111/j.2044-8295.1991.tb02384.x.

Valentine, T., Lewis, M.B. and Hills, P.J. (2016) 'Face-Space: A Unifying Concept in Face Recognition Research', *Quarterly Journal of Experimental Psychology*, 69(10), pp. 1996–2019. Available at: https://doi.org/10.1080/17470218.2014.990392.

Van Duynslaeger, M. *et al.* (2008) 'EEG components of spontaneous trait inferences', *Social Neuroscience*, 3(2), pp. 164–177. Available at: https://doi.org/10.1080/17470910801907226.

Van Duynslaeger, M., Van Overwalle, F. and Verstraeten, E. (2007) 'Electrophysiological time course and brain areas of spontaneous and intentional trait inferences', *Social Cognitive and Affective Neuroscience*, 2(3), pp. 174–188. Available at: https://doi.org/10.1093/scan/nsm016.

Van Overwalle, F. (2009) 'Social cognition and the brain: A meta-analysis', *Human Brain Mapping*, 30(3), pp. 829–858. Available at: https://doi.org/10.1002/hbm.20547.

Varnum, M.E.W. *et al.* (2012) 'Social class differences in N400 indicate differences in spontaneous trait inference', *Journal of Experimental Psychology: General*, 141(3), pp. 518–526. Available at: https://doi.org/10.1037/a0026104.

Verfaellie, M. (2002) 'The effect of retrieval instructions on false recognition: exploring the nature of the gist memory impairment in amnesia', *Neuropsychologia*, 40(13), pp. 2360–2368. Available at: https://doi.org/10.1016/S0028-3932(02)00074-X.

Vetter, B. (2015) *Potentiality: From Dispositions to Modality*. Oxford University Press.

de Vignemont, F. (2009) 'Drawing the boundary between low-level and high-level mindreading', *Philosophical Studies*, 144(3), pp. 457–466. Available at: https://doi.org/10.1007/s11098-009-9354-1.

Vugt, M.V. and Grabo, A.E. (2015) 'The Many Faces of Leadership: An Evolutionary-Psychology Approach', *Current Directions in Psychological Science* [Preprint]. Available at: https://doi.org/10.1177/0963721415601971.

Vukasović, T. and Bratko, D. (2015) 'Heritability of personality: A meta-analysis of behavior genetic studies.', *Psychological Bulletin*, 141(4), pp. 769–785. Available at: https://doi.org/10.1037/bul0000017.

Wahl, O. (1932) 'Neue Untersuchungen über das Zeitgedächtnis der Bienen. [Further investigations on the temporal memory of bees.]', *Zeitschrift für Vergleichende Physiologie*, 16, pp. 529–589.

Wang, S.-H., Baillargeon, R. and Brueckner, L. (2004) 'Young infants' reasoning about hidden objects: evidence from violation-of-expectation tasks with test trials only', *Cognition*, 93(3), pp. 167–198. Available at: https://doi.org/10.1016/j.cognition.2003.09.012.

Wei, Y., Schatten, H. and Sun, Q.-Y. (2015) 'Environmental epigenetic inheritance through gametes and implications for human reproduction', *Human Reproduction Update*, 21(2), pp. 194–208. Available at: https://doi.org/10.1093/humupd/dmu061.

Wellman, H.M., Cross, D. and Watson, J. (2001) 'Meta-analysis of theory-of-mind development: the truth about false belief', *Child Development*, 72(3), pp. 655–684.

Westra, E. (2018) 'Character and theory of mind: an integrative approach', *Philosophical Studies*, 175(5), pp. 1217–1241. Available at: https://doi.org/10.1007/s11098-017-0908-3.

Westra, E. (2019) 'Getting to Know You: Accuracy and Error in Judgments of Character', *Mind and Language* [Preprint].

Westra, E. (2020) 'Getting to know you: Accuracy and error in judgments of character', *Mind & Language*, 35(5), pp. 583–600. Available at: https://doi.org/10.1111/mila.12258.

Westra, E. (2021) 'Folk personality psychology: mindreading and mindshaping in trait attribution', *Synthese*, 198(9), pp. 8213–8232. Available at: https://doi.org/10.1007/s11229-020-02566-7.

Wheeler, M. (2017) 'The Revolution will not be Optimised: Radical Enactivism, Extended Functionalism and the Extensive Mind', *Topoi*, 36(3), pp. 457–472. Available at: https://doi.org/10.1007/s11245-015-9356-x.

Wicker, B. *et al.* (2003) 'Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust', *Neuron*, 40(3), pp. 655–664. Available at: https://doi.org/10.1016/S0896-6273(03)00679-2.

Wille, B., De Fruyt, F. and De Clercq, B. (2013) 'Expanding and Reconceptualizing Aberrant Personality at Work: Validity of Five-Factor Model Aberrant Personality Tendencies to Predict Career Outcomes', *Personnel Psychology*, 66(1), pp. 173–223. Available at: https://doi.org/10.1111/peps.12016.

Willis, J. and Todorov, A. (2006) 'First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face', *Psychological Science*, 17(7), pp. 592–598. Available at: https://doi.org/10.1111/j.1467-9280.2006.01750.x.

Wimmer, H. and Perner, J. (1983) 'Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception', *Cognition*, 13, pp. 103–128. Available at: https://doi.org/10.1016/0010-0277(83)90004-5.

Winawer, J. *et al.* (2007) 'Russian blues reveal effects of language on color discrimination', *Proceedings of the National Academy of Sciences of the United States of America*, 104(19), pp. 7780–7785. Available at: https://doi.org/10.1073/pnas.0701644104.

Winter, L. and Uleman, J.S. (1984) 'When are social judgments made? Evidence for the spontaneousness of trait inferences', *Journal of Personality and Social Psychology*, 47(2), pp. 237–252. Available at: https://doi.org/10.1037/0022-3514.47.2.237.

Witt, C. (2003) 'The Priority of Actuality', in *Ways of Being*. Cornell University Press (Potentiality and Actuality in Aristotle's Metaphysics), pp. 75–96. Available at: http://www.jstor.org/stable/10.7591/j.ctv1fxmvt.8 (Accessed: 9 February 2022).

Witzel, C. and Gegenfurtner, K.R. (2015) 'Categorical facilitation with equally discriminable colors', *Journal of Vision*, 15(8), p. 22. Available at: https://doi.org/10.1167/15.8.22.

Witzel, C. and Gegenfurtner, K.R. (2018) 'Color Perception: Objects, Constancy, and Categories', *Annual Review of Vision Science*, 4(1), pp. 475–499. Available at: https://doi.org/10.1146/annurev-vision-091517-034231.

Wright, B. *et al.* (2016) 'Social Stories™ to alleviate challenging behaviour and social difficulties exhibited by children with autism spectrum disorder in mainstream schools: design of a manualised training toolkit and feasibility study for a cluster randomised controlled trial with nested qualitative and cost-effectiveness components', *Health Technology Assessment*, 20(6), pp. 1–258. Available at: https://doi.org/10.3310/hta20060.

Xiao, Y. *et al.* (2011) 'The Biological Basis of a Universal Constraint on Color Naming: Cone Contrasts and the Two-Way Categorization of Colors', *PLOS ONE*, 6(9), p. e24994. Available at: https://doi.org/10.1371/journal.pone.0024994.

Yuill, N. (1997) 'Children's understanding of traits', in *The development of social cognition*. Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis (Studies in developmental psychology), pp. 273–295.

Yuill, N. and Pearson, A. (1998) 'The development of bases for trait attribution: Children's understanding of traits as causal mechanisms based on desire.', *Developmental Psychology*, 34(3), pp. 574–586. Available at: https://doi.org/10.1037/0012-1649.34.3.574.