ABSTRACT

Title of Dissertation:	THE ARCHITECTURE AND DEVELOPMENT OF MINDREADING: BELIEFS, PERSPECTIVES, AND CHARACTER
	Evan Edward Westra, Doctor of Philosophy, 2017.
Dissertation directed by:	Professor Peter Carruthers, Department of Philosophy

This dissertation puts forward a series of arguments and theoretical proposals about the architecture and development of the human capacity to reason about the internal, psychological causes of behavior, known as "theory of mind" or "mindreading."

Chapter 1, "Foundations and motivations," begins by articulating the philosophical underpinnings of contemporary theory-of-mind debates, especially the dispute between empiricists and nativists. I then argue for a nativist approach to theory-of-mind development, and then go on to outline how the subsequent chapters each address specific challenges for this nativist perspective.

Chapter 2, "Pragmatic development and the false-belief task," addresses the central puzzle of the theory-of-mind development literature: why is it that children below the age of five fail standard false-belief tasks, and yet are able to pass implicit versions of the false-belief task at a far younger age? According to my novel, nativist

account, while they possess the concept of BELIEF very early in development, children's early experiences with the pragmatics of belief discourse initially distort the way they interpret standard false-belief tasks; as children gain the relevant experience from their social and linguistic environment, this distortion eventually dissipates. In the Appendix (co-authored with Peter Carruthers), I expand upon this proposal to show how it can also account for another set of phenomena typically cited as evidence against nativism: the Theory-of-Mind Scale.

Chapter 3, "Spontaneous mindreading: A problem for the two-systems account," challenges the "two-systems" account of mindreading, which provides a different explanation for the implicit/explicit false-belief task gap, and has implications for the architecture of mature, adult mindreading. Using evidence from adults' perspective-taking abilities I argue that this account is theoretically and empirically unsound.

Chapter 4, "Character and theory of mind: An integrative approach," begins by noting that contemporary accounts of mindreading neglect to account for the role of character or personality-trait representations in action-prediction and interpretation. Employing a hierarchical, predictive coding approach, I propose that character-trait representations are rapidly inferred in order to inform and constrain our mental-state attributions.

Because this is a "covering concept" dissertation, each of these chapters (including the Appendix) is written so that it is independent of all of the others; they can be read in any order, and do not presuppose one another.

THE ARCHITECTURE AND DEVELOPMENT OF MINDREADING: BELIEFS, PERSPECTIVES, AND CHARACTER

by

Evan Edward Westra

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Peter Carruthers, Chair
Professor Georges Rey
Associate Professor Alexander Williams
Assistant Professor Jonathan Beier
Professor Jeffrey Lidz

Dedication

For Monique, Haijo, and Adam.

Acknowledgements

I am deeply indebted to a great many people who have supported me throughout my time writing this dissertation, and throughout my academic career more generally. My greatest debt is undoubtedly to the members of my immediate family, to whom this dissertation is dedicated. My mother Monique Westra has been a source of constant love and encouragement during the times when I have struggled. My brother Adam Westra, in whose footsteps I followed to become a philosopher, has not only been a source of inspiration for me, but also my best friend. I am most indebted to my father Haijo Westra, who has read every word that I have ever written. His intellect, patience, and generosity are virtues that I will forever aspire to match. I am also profoundly very grateful to my partner, Laura Elenbaas, for her intellectual and emotional companionship throughout our time at Maryland and beyond.

My other great debt is to my supervisor, Peter Carruthers. Peter is an incredible, prolific philosopher whose work has profoundly shaped my own; but he is also a kind, diligent, and incomparably generous mentor. I am extremely lucky to have been his student. Many thanks are also owed to Jonathan Beier and Georges Rey, who have both played important roles in my development as a scholar.

I am fortunate to have received support and comments on my work from a number of other faculty, friends, and colleagues: Rachel Dudley, Valentine Hacquard, Joseph Jebari, Andrew Knoll, John Michael, Paul Pietroski, Brendan Ritchie, Julius Schönherr, Shannon Spaulding, Moonyoung Song, Charles Starkey, J. Robert Thompson, and Alexander Williams. Special thanks are due to Brandon Terrizzi, who has been immensely helpful to me as a colleague, a collaborator, and a friend. I am

also grateful for comments from anonymous reviewers and editors from the following journals: *Cognition, Philosophical Studies, Review of Philosophy and Psychology,* and *Synthese*.

This research was supported by the Social Sciences and Humanities Research Council of Canada, doctoral fellowship #752-2015-0035.

Table of Contents

Chapter 1: Foundations and motivations

1. Introduction

This dissertation is comprised of three core papers about theory of mind, or "mindreading," which is the capacity to represent the mental states of other agents for the purposes of predicting and interpreting behavior. Each paper is concerned with a distinct aspect of theory of mind: Chapter 2 focuses on a longstanding debate about when we acquire the concept of BELIEF; Chapter 3 uses data on our reasoning about perceptual states to make a point about the architecture of theory of mind; Chapter 4 proposes a model of how we reason about character traits. Additionally, I include in the Appendix a follow-up to the account laid out in Chapter 2, written collaboratively with Peter Carruthers, which concerns children's ability to reason about desires and states of knowledge. The three core chapters are independent of one another, and can be read in any order. However, these three papers share a common approach, and can be read as part of a broader project. This project begins with a nativist perspective on the acquisition of theory of mind, and then works through a succession of key challenges from the mindreading literature that such an account must overcome. As these challenges are addressed one by one, a broader picture of the architecture and development of theory of mind starts to emerge.

To establish this way of reading this dissertation, I begin by discussing the dispute between nativists and empiricists in the abstract, and laying out the general motivations for adopting a nativist approach. Then, I introduce a debate in the

mindreading literature that hovers in the background of all three chapters in my dissertation: whether or not infants have the capacity to represent beliefs. This subject has been a source of renewed controversy among philosophers and developmental psychologists for over a decade now, epitomizing the contrast between nativist and empiricist approaches to mindreading, and to the mind more generally. Having established the plausibility of a nativist approach to theory of mind, I go on to present a few of the key empirical challenges that this view faces, and summarize how they are addressed in the chapters to come.

2. What's the philosopher doing here?

First, however, a few words about why I, as a philosopher, am writing about this topic at all. At a glance, the topic of mindreading seems to belong to the domain of psychology. As such, one might wonder what a philosopher is doing writing about it. Fortunately for me, I am not alone: since it became a topic of sustained empirical research, philosophers have managed to insert themselves into debates about mindreading. When Premack and Woodruff, in their seminal article in *Brain and Behavioral Sciences*, asked, "Does the Chimpanzee have a Theory of Mind?", it was three philosophers – Dennett, Harman and Bennett – who, in independent commentaries, suggested a principled criterion for discovering the answer: show that chimpanzees understand the *intentionality* of mental states (Bennett, 1978; Dennett, 1978; Harman, 1978; Premack & Woodruff, 1978). A good way to test this, they speculated, would be to design a behavioral task that could only be solved correctly given an understanding of *false* beliefs. A few years later, Wimmer and Perner (1983) implemented the first-false belief task with young children, initiating a massive

developmental research program that continues to this day. During this time, philosophers have continued to play a substantial role in the mindreading literature (Butterfill & Apperly, 2013; Carruthers, 2013; Christensen & Michael, 2015; Fenici, 2013; Fodor, 1992; Goldman, 2006; Gordon, 1986; Heal, 1996; Nichols & Stich, 2003; Spaulding, 2010; Thompson, 2014; Zawidzki, 2011).

One of the reasons why philosophers have been so involved in the mindreading debate has to do with its similarity to an old philosophical problem. Arising from the Cartesian thesis that we have privileged access to our own minds, and can therefore be certain of our own existence, the problem of other minds points to the fact that this certainty does not extend to the existence of other minds. Moreover, our understanding of our own mind seems to be intrinsically bound to our own first-person experience. Thus, there is a profound conceptual and epistemological gap between our grasp of our own minds and the minds of others. Given this gap, it is not clear how we could ever conceive of a mind that is not our own, let alone have knowledge of such a thing (Avramides, 2000; Hyslop, 2014).

In answer, the philosophical literature has produced a number of attempted solutions. Mill proposed that our knowledge of other minds is grounded in an analogy between our own experience of the connection between our first-person mental states and behaviors, and our observations of the behaviors of others (Mill, 1865). Another solution is that we posit the existence of other minds because this is what best explains our observations of regularities in behavior; thus, other minds are treated as theoretical posits, much like other unobservables in scientific theories (Pargetter, 1984). Still other philosophers have attempted to deflate the problem, either by

endorsing a form of behaviorism about mental states (Wittgenstein, 1953), or by claiming perception-like access to some of other agents' mental states (Cassam, 2007; Gallagher, 2008).

There is both a "hard" and a "soft" problem of other minds. The "hard" problem concerns how we reason about or conceptualize the qualitative, phenomenally conscious aspects of other minds: how do we know that other people's experiences are like ours, or completely different? How do we know that other people are not really philosophical zombies, who are functionally and physically like us, but lack inner lives and subjective experiences? The "soft" problem of other minds, on the other hand, assumes a functionalist account of the mind, and does not explicitly concern itself with phenomenal consciousness or qualia. It simply asks us how it is that we conceive of and reason about the existence of unobservable, intentional states that interact to cause behavior.

The hard problem, like the hard problem of consciousness, is widely believed to be intractable. The soft problem, on the other hand, really amounts to an empirical puzzle about the nature of human reasoning about other agents. And while the analogical and (especially) theory-based strategies for solving the hard problem seem to fall short of what would be needed to overcome the gap between first- and third-person experiences of other minds, they offer plausible psychological hypotheses for solving the soft problem. As a result, it has been the soft problem of other minds, and not the hard problem, that has really influenced the contemporary mindreading debate. The debate between the analogical and theoretical solutions to the problem of other minds gave way to the simulation theory/theory-theory debate in the

contemporary theory-of-mind literature. Today, most mindreading theorists identify their views with some form of theory-theory, simulation theory, or a hybrid of the two (Carruthers & Smith, 1996).

This is a telling bit of intellectual history: what began as a single, seemingly insoluble philosophical problem came to structure the solution-space for an entire sub-field of empirical research. And this influence of philosophy on scientific practice was not merely historical: as these philosophical frameworks were put into practice, philosophers continued to refine and update them in light of new evidence, leading to increasingly sophisticated psychological models of mindreading (incorporating, for example, insights from dual-process approaches to cognition (Goldman, 2006)). We see something similar in the way philosophers contributed to the creation of the false-belief task: initially, philosophers established the conceptual criteria for establishing an understanding of other minds, and then psychologists went ahead and put it into practice. As time has passed, philosophers have elaborated upon these conceptual criteria, and psychologists have developed new experimental designs as a result. To take a notable example, psychologists have now taken a cue from Frege, and developed tasks that test children's grasp of the fact that two people can represent the same object under different modes of presentation (Low, Drummond, Walmsley, & Wang, 2014; Rakoczy, 2015). Increasingly often, this combination of philosophical ideas and psychological practice has come in the form of direct collaborations between members of both disciplines (Butterfill & Apperly, 2013).

It is not an exaggeration to say that the study of mindreading has become the paradigm of cooperation between philosophers and scientists. This is no accident: the

discipline of philosophy is a deep well of concepts and arguments that stand to illuminate our scientific understanding of the world. As scientists design experiments, develop new ontologies, and choose between competing explanations, they often draw from this well, and start to engage in philosophy themselves. Simultaneously, we philosophers often find ourselves drawn from more abstract questions towards the activities of scientists. At first, this happens because we think empirical data might inform our more traditional, philosophical debates. But then we notice something: the scientists are doing philosophy as well, on questions just as profound as any that we concern ourselves with! Thus, philosophers and scientists come to meet each other in the middle, discovering that they possess complementary skills and strengths. The convergence of philosophers and scientists has the potential to generate profound contributions to our understanding of the world. This potential is on full display in the mindreading literature. My own research falls within this tradition.

3. From empiricism to nativism

While the problem of other minds has significantly shaped the contemporary mindreading literature, some of the deepest disagreements in this field stem from the dispute between empiricism and nativism. While the historical empiricist-nativist debate was often also concerned with epistemological issues, the contemporary one is more narrowly focused on the kinds of psychological representations and processes that support learning. Generally speaking, empiricists explain learning as the product of domain-general inferential processes, perceptually based representations, and a very small set of innately specified primitives. Nativists, in contrast, posit additional domain-specific inferential processes and innate representations. The disagreement

between nativists and empiricists thus comes down to a dispute about the extent of children's innate learning endowment – what Margolis and Laurence call the "acquisition base" (Margolis & Laurence, 2012). In this section, I present some very general considerations in favor of adopting a nativist approach, even if we take the existence of domain-general learning systems as a theoretical point of departure.

To better understand the nature of the nativist-empiricist dispute, it is useful to consider how it manifests itself in discussions of Bayesian models of cognitive development, which are currently in vogue in empiricist circles (Gopnik & Wellman, 2012; Perfors, Tenenbaum, Griffiths, & Xu, 2011; Tenenbaum et al., 2011). Learning, according to this view, basically amounts to a form of Bayesian hypothesis testing, whereby the posterior probability of a hypothesis given a particular observation is derived from prior probabilities assigned to hypothesis and the observation, as well as the likelihood of the observation given the hypothesis. Bayesian models of cognitive development assume that children learn to use their observations and interactions with the world to update empirical hypotheses about the causes of their sensory experiences. More sophisticated "hierarchical" Bayesian models of cognitive development suggest that children do this at multiple levels of abstraction simultaneously, updating "overhypotheses" about the relative probabilities of different hypothesis spaces at the subordinate level (Kemp, Perfors, & Tenenbaum, 2007). Collectively, this hierarchy of Bayesian updating procedures is said to provide a powerful computational framework for learning about the world. Note that *qua* learning procedure, Bayesian hypothesis testing is totally domain-general, and can be

applied to any set of hypotheses and observations, regardless of their content – a feature that makes these models attractive to empiricists.

One might wonder how such a model initially assigns prior probabilities to different hypotheses, or if it starts out with a flat probability distribution. But even before we can ask this question, we must consider where the hypotheses themselves come from. Notably, Bayesian models of learning do not actually provide an explanation for how hypotheses are generated in the first place, or how the space of hypotheses gets expanded over time. The Bayesian formalism only gives the procedure for updating the probabilities of pre-existing hypotheses. Thus, the acquisition base for Bayesian models of cognitive development must include additional, constructive procedures that are capable of generating new hypotheses, as well as the representational primitives that are involved in the hypothesis-construction-process. Following Perfors and colleagues, let us call the entire range of hypotheses that these procedures are capable of generating the *latent hypothesis space*, and the hypotheses that are actually subject to testing and updating the *explicit hypothesis space*. The process of hypothesis generation thus describes the process whereby hypotheses are moved from the latent space to the explicit one (Perfors et al., 2011).

Importantly, the procedures and representational primitives contained in the latent hypothesis space would necessarily determine the range of explicit hypotheses that a Bayesian learner could potentially entertain. A latent acquisition base lacking the capacity to represent SQUARE, for instance, could never generate hypotheses

about squares, which would then make learning about squares impossible. It would also be entirely consistent with the Bayesian approach if the hypothesis-construction process contained *every* possible concept as a representational primitive. Indeed, Fodor famously envisioned a hypothesis-testing account of learning that included all possible concepts in the acquisition base, including GRANDMOTHER and DOORKNOB (Fodor, 1975). No doubt this particular proposal is not what empiricists have in mind when they invoke Bayesian models of cognitive development. But this only serves to highlight the fact that Bayesian hypothesis-testing accounts of learning are not uniquely empiricist: indeed, they are compatible with even the most radical forms of nativism.

Thus, even if we assume that a Bayesian hypothesis-testing account is correct, the true mark of an empiricist model of learning will not be a domain-general learning procedure as such, but rather a domain-general approach to hypothesis construction. For traditional empiricist accounts, this will consist in a limited set of perceptually based representational primitives, associative procedures that track statistical relations between different combinations of these primitives, and abstraction procedures that derive more general representations (e.g. concepts of categories) from particular perceptual inputs (Hume, 2000; Locke, 1690; see also Hornstein, 2005). With this kind of acquisition base, new explicit hypotheses about present and future inputs can be generated based on representations of past inputs alone. However, a crucial drawback of such a hypothesis-construction procedure is that the latent hypothesis space will be limited by the particular inputs that it has already processed. For a machine-learning algorithm that only needs to construct hypotheses about a particular

domain, and is fed a large number of inputs related to that domain, this can work fine. But for a human infant, who must use her limited experiences in order to learn about many different domains, this mode of hypothesis construction is inevitably inflexible.

More sophisticated empiricist models might enrich this set of tools with the capacity to construct totally novel hypotheses that have never been experienced at all. One example of this kind of enriched empiricist framework is Gopnik and colleagues' proposal that learners generate structured representations of possible causal patterns in the environment, or "causal Bayesian networks" (Gopnik & Schulz, 2004; Pearl, 2000). Causal Bayesian networks go beyond simple associative models in several ways. First, they can represent different relations between co-occurring representations, which are treated as variables: whereas an associative model can only represent that A, B, and C co-occur, causal models can represent A as the cause of B and B as the cause of C, or A as the common cause of both B and C, and so on. Further, for each of these different causal representations, the network encodes corollary information about how changes to one of the causal variables would affect the others (e.g. if A causes B and B causes C, a change to B will imply a change in C; but if A is the common cause of B and C, then changes to B would leave C unaffected). Causal Bayesian networks are also capable of introducing and inferring *hidden variables*, which enables them to form hypotheses about causal structures whose component elements are not directly observable. This account of hypothesis formation has a much wider range than basic association-based models: on this view, every permutation and combination of causal relations between variables can be expressed as a testable hypothesis about the world. Notably, this flexibility was only

achieved by adding knowledge of causality to the acquisition base – in effect, making the account more nativist.

However, this added flexibility also creates new challenges. If the latent hypothesis space now contains every possible causal pattern that could obtain between a set of variables, this means that the space of possible hypotheses will now be extremely large (if not infinite). This raises the question of which hypotheses the learner should generate and test *first*. If the learner were to begin by assuming a flat probability distribution across the entire hypothesis space, then it would have no choice but to begin generating and testing hypotheses at random. Given the size of the latent hypothesis space, this makes for a highly inefficient strategy: a learner could spend all its time generating and testing completely spurious hypotheses, without ever actually learning anything about the world. Thus, equipping a Bayesian learner with a highly flexible hypothesis-generation procedure creates a learning problem that is the inverse of the associationist hypothesis-generation strategy: if the latent hypothesis space is too large, then learning becomes extremely difficult.

The solution to this problem is to impose some structure on the hypothesis-generation procedure, so that it is no longer randomly generating and testing individual hypotheses. One way to do this is to posit overhypotheses about which portions of the latent hypothesis space will yield fruitful hypotheses – essentially, a hypothesis about a prior probability distribution over the whole hypothesis space. A Bayesian learner could then start by testing overhypotheses, and thereby arrive at a non-flat probability distribution over space of possible hypotheses, which would then guide the generation and testing of hypotheses. However, this strategy risks running

into the same problem as before: if the space of overhypotheses is also extremely large and has a flat probability distribution, then a learner will have no choice but to start generating and testing overhypotheses at random. Thus, it may be just as difficult to arrive at informative overhypotheses as it is to arrive at informative first-order hypotheses.

Ultimately, the only way to resolve this issue is to posit that the learner starts out biased towards certain overhypotheses, and that these enable her to start generating and testing first-order hypotheses that will actually be informative. With such initial biases in place, a learner could thus avoid getting stuck in a fruitless pattern of random hypothesis-generation and testing, and instead actually learn things about the world. However, these biases could not themselves be the product of the same hypotheses-generation and testing procedure that they are intended to constrain. If they were, then we would be faced with the problem of explaining how we arrived at *those particular* biases as opposed to other possible ones. Thus, the biases that initially constrain the learning procedure must not themselves be the product of a learning procedure – in other words, they must be innate.

And so, even if we start with an empiricist approach to learning, and allow for empiricist hypothesis-generation mechanisms, we still need to posit additional innate constraints in order to make the learning process tractable. This means that the nativist might be in complete agreement about the idea that learning consists in a domain-general Bayesian updating procedure that tracks statistical regularities in the environment, and even the idea that we possess domain-general hypothesis-construction procedures. But this cannot be the whole story. Nativists, on this picture

of the empiricist-nativist dispute, are simply interested in describing those additional structures, how they relate to particular kinds of hypotheses about the world, and the ways in which they facilitate learning. Empiricists, in contrast, are interested in describing the basic domain-general structures that apply to all forms of hypothesis testing. The disagreement between these two camps ultimately comes down to a question about the extent of these innate constraints on learning. Empiricists think that we should posit fewer innate constraints on learning, while nativists are open to the possibility that these constraints might be quite extensive.

4. Nativism, modularity, and cognitive architecture

Until this point, we've been working with the abstract assumption that hypothesis-generation involves some combination of perceptually based representational primitives, abstraction, and domain-general causal modeling. But one way that innate constraints on learning might manifest themselves is via innately channeled mechanisms that spontaneously generate particular kinds of hypotheses about the world in response to environmental input. If, for instance, learners spontaneously interpret self-generated movements as caused by *intentional agents* with *goals* and *perceptual states*, then the hypotheses they generate in response to perceptions of self-generated movement will tend to invoke intentional concepts (Gergely & Csibra, 2003). Likewise, learners might spontaneously interpret sounds coming from agents as linguistic utterances, and seek to extract from them syntactic structure and semantic content that conforms to the parameters of Universal Grammar. These are examples of domain-specific hypothesis-generating systems that could lead learners to apply innate concepts to their experiences. Contemporary nativist proposals have

mostly been about the existence of such domain-specific conceptual systems and their particular attributes.

To cash out these proposals in architectural terms, many nativists have appealed to the concept of modularity. However, the concept of modularity comes in varying strengths, a fact that often leads to confusion among critics of nativism. The strongest conception of modularity was originally invoked by Fodor to describe the innate architecture underlying perceptual and linguistic processing (Fodor, 1983). According to this view, modules tend to possess nine features: domain-specificity, mandatory operation, limited central accessibility, fast processing, informational encapsulation, 'shallow' outputs, fixed neural architecture, characteristic and specific breakdown patterns, and innateness. However, even Fodor noted that a cognitive system could count as modular even if it adhered to these properties only "to some interesting extent," suggesting that he did not see all of these nine features as necessary for modularity; indeed, there are virtually no proponents of modularity today who endorse all of the items on this list as necessary and sufficient conditions.

One of the stronger approaches to modularity takes informational encapsulation and limited central accessibility as its signature properties (Apperly, 2011; Coltheart, 1999; Scholl & Gao, 2013; Scholl & Leslie, 1999). According to this approach, modules constrain hypothesis-generation by applying a specific set of computational procedures to specific kinds of perceptual inputs. In generating its output, the module is only able to use information carried by the input signal, and information stored internally within the module itself. It cannot draw on any information stored outside of the module. Conversely, the module-internal processing

is also inaccessible to other computational processes taking place outside the module, including person-level goals and intentions. This ensures that given a particular set of environmental inputs, modules are guaranteed to produce a particular set of outputs. Once an output is produced, it can either serve as input into another module, or figure in domain-general inferences.

As a consequence of these informational restrictions, the module exhibits a number of the other features of modularity from Fodor's original list: because it is only sensitive to certain inputs, it is domain-specific; because its operations are not accessible to goals, it is mandatory; because it only draws on information internal to the module, it is fast and efficient; because it cannot draw on information from outside the module, its outputs are representationally "shallow." One of the features of modularity that is not implied by this account is innateness, however: it is quite plausible that encapsulated systems could emerge through a process of environmentally-driven specialization, or "downwards modularization" (d'Souza & Karmiloff-Smith, 2011). Encapsulation-based views of modularity thus make fairly specific claims about the how a system processes information, but need not imply that that system has an innate basis.

In contrast to the encapsulation-based approach, which preserves the bulk of the features associated with Fodor's original account of modularity, some conceptions of modularity are quite weak, and only posit that modules are domain-specific computational systems (Barrett & Kurzban, 2006; Carruthers, 2006). On this view, to postulate the existence of a module is basically to posit that a particular cognitive function is somehow performed by the brain. In the context of nativism, the relevant

sense of "function" invoked by this conception of modularity is often tied to natural selection: to say that an organism possesses an innate module that performs some function F is to say that doing F was selected for in an ancestral population. Thus, while weak modularity makes few specific claims about cognitive architecture, it does tend to imply an evolutionary argument for why we should expect a particular function to be innate in the first place.

Despite the association between nativism and modularity, there is no obvious theoretical reason to think that innate learning constraints should be modular in the strong sense. Some innate structures may be rigidly encapsulated, but it is entirely possible that many innate structures are modular only in the weaker, functional sense. Moreover, some innate systems may include components that are modular in *both* the strong *and* weak senses. Carey, for instance, argues that the core knowledge system for object knowledge includes both a perceptual component that is modular in the strong sense, but that the outputs of this module – domain-specific, specialized representations for object-tracking called "object files" – can be "bound" with content from outside the module, and can enter into domain-general inferences (Carey, 2009). Other nativists have posited innately channeled systems operating on a few core principles that initially display a kind of encapsulation due to limited processing resources and experience; however, as children mature, these systems become enriched over time with new principles derived from experience (Baillargeon, 2008). The point is that positing innate constraints on learning in a particular domain need not imply any particular architectural commitment. Which cognitive structures

underlie innate structures in a given domain is an empirical issue, and could very well vary on a case-by-case basis.

5. Empiricism, nativism, and the false-belief controversy

In the developmental literature on the topic, empiricist-nativist debates about mindreading have proceeded in a relatively piecemeal fashion. Instead of taking mindreading capacities as a monolithic whole, developmental psychologists have instead investigated the sources of particular concepts of mental states, including beliefs, goals, intentions, emotions, and perceptual states. This stands to reason: just as one need not be a nativist or empiricist about knowledge as a whole, one need not be a nativist or empiricist about every aspect of theory of mind. For instance, concepts of states like embarrassment or guilt – complex compounds of beliefs and emotions that get tokened in response to the violation of social norms – seem unlikely to be innate. Concepts of mental states with modal contents, such as *supposition* and *wondering*, also seem to be quite complex, and few nativists would posit that they are innate. On the other hand, many are inclined to accept that very young children possess a basic grasp of perceptual and teleological relations that form the rudiments of concepts like *goal* and *sees* (Butterfill & Apperly, 2013; Carey, 2009; Gergely & Csibra, 2003).

In principle, it is possible that innate mental-state understanding could be atomistic: an agent could possess one mental-state concept (e.g. SEEING), but lack all of the others. As an empirical hypothesis, however, this seems unlikely. If, as many theorists hold, the function of mindreading is the prediction and interpretation of behavior, then mindreaders must be able to grasp the relations between mental

states and behavior. This requires an understanding of practical reasoning, which implies a grasp of the inferential relations between beliefs and desires. And so, if the thesis that we possess any mental-state concepts innately is meant to explain how we come to predict and interpret behavior, then this thesis must hold that we possess multiple mental-state concepts. It must also hold that we understand something about the inferential relations between these concepts, and something about the causal relations between practical reasoning and action. Thus, while empirical research on mindreading might proceed on a concept-by-concept basis, nativist accounts of mindreading must posit a cohesive set of mindreading abilities and concepts.

That said, a great deal of the empiricist-rationalist debate in the mindreading literature has focused on whether our innate set of mindreading abilities includes the ability to represent hypotheses about *beliefs*. Beliefs are thought to be special because they are paradigmatic bearers of intentionality, a property widely held to be the hallmark of the mental (Chisholm, 1967). Beliefs are robustly intentional: they can be false, they can be about entities that do not exist, and they are inherently aspectual. In contrast, some theorists have claimed that infants might initially possess non-intentional versions of the concepts SEEING or GOAL that refer to simple spatial relations between agents and the environment (Flavell, Everett, Croft, & Flavell, 1981; Gergely & Csibra, 2003). These relations are not genuinely intentional, because they can only hold between agents and real properties of the world, and therefore do not permit misrepresentation. Thus, while a grasp of perceptual and teleological states need only imply a rudimentary grasp of other minds, understanding beliefs means

understanding intentionality, and thus implies the possession of a representational theory of mind.

Since the early 1980s, the gold standard for measuring children's understanding of beliefs has been the false-belief task (Wimmer & Perner, 1983). In the classic Sally-Anne version of this task (Baron-Cohen, Leslie, & Frith, 1985), children are shown a vignette in which one character (Sally), places a marble in a basket, and then leaves the room. While she is gone, another character (Anne) moves the marble from the basket to a box. When Sally returns, the experimenter asks the child where Sally will look for the marble. The correct answer, of course, is that Sally will look for the marble in the basket, because that is where she falsely believes it to be. Strikingly, young children are overwhelmingly likely to systematically fail this task, and say that Sally will look for the marble in the box, its actual location.

Until the early 2000s, the broad consensus among developmental researchers was that this showed that children acquire the concept of BELIEF around four years of age (Wellman, Cross, & Watson, 2001). Typically, the developmental explanations for this invoke a number of experiential factors, including exposure to mental state terms (Dunn & Brophy, 2005), propositional-embedding syntax (de Villiers & Pyers, 2002), and various forms of social experience (Tomasello & Rakoczy, 2003). Some of these explanations situate the acquisition of a representational concept of BELIEF within a more extended, experience-driven developmental progression through different stages of mental-state understanding (Gopnik & Wellman, 1992; Wellman & Liu, 2004). What all of these theories have in common is that they posit that a combination of domain-general learning and some specific set of environmental

factors ultimately explain the emergence of false-belief understanding around four years of age. In other words, the consensus view was that some kind of empiricist model would explain the acquisition of BELIEF.

However, a small number of philosophers and psychologists have consistently questioned this empiricist consensus. The standard, verbal false-belief task, they argued, fails to get at children's underlying competence with the concept of BELIEF, since, in addition to an understanding of beliefs, it requires advanced linguistic and executive abilities (P. Bloom & German, 2000; Fodor, 1992; Scholl & Leslie, 1999, 2001). Pockets of evidence seemed to support these worries. The finding that children with autism tended to have difficulty with the false-belief task even as they performed well on comparable tasks led some researchers to postulate that theory of mind is subserved by an innately channeled modular architecture (in the strong, encapsulation-based sense) (Baron-Cohen et al., 1985; Baron-Cohen, 1997; Leslie, Friedman, & German, 2004). This approach led these researchers to discover that experimental manipulations aimed at easing the attentional demands of the false belief task could enable children to pass it earlier than previously thought (Siegal & Beattie, 1991; Surian & Leslie, 1999). Even proponents of the received view occasionally ran into anomalous findings: Clements and Perner discovered that younger children seem to demonstrate an "implicit" understanding of beliefs through their looking behavior even as they failed verbal false belief tasks (Clements & Perner, 1994).

Motivated by these considerations, as well as a growing literature on the considerable social-cognitive abilities of very young infants (Gergely & Csibra, 2003;

Woodward, 1998), developmental researchers began to apply non-verbal methods designed to spontaneously elicit infant knowledge to test younger children's abilities to attribute false beliefs. The results of these first studies turned the received view of theory of mind development on its head, apparently vindicating the nativist minority: Onishi and Baillargeon's seminal study (2005) used a looking time measure to show that 15-month old infants form expectations about behavior that are sensitive to agents' false beliefs. In the past ten years, this finding has been replicated dozens of times using a wide range of methods, and evidence of false belief understanding has been demonstrated in infants as young as six months of age (Baillargeon, Scott, & He, 2010; Barrett et al., 2013; D. Buttelmann, Carpenter, & Tomasello, 2009; D. Buttelmann, Over, Carpenter, & Tomasello, 2014; Kovács, Téglás, & Endress, 2010; Senju, Southgate, Snape, Leonard, & Csibra, 2011; Southgate & Vernetti, 2014).

Empiricists about belief understanding have responded by offering deflationary, non-mentalistic interpretations of the infant false-belief data. They argue that infants might pass these tasks even without the capacity to represent beliefs; instead, infants' expectations in these tasks might simply be driven by low-level statistical associations formed either during their everyday experiences (Perner, 2010) or during the task itself (Heyes, 2014a; Ruffman, 2014). Defenders of belief-nativism have addressed these arguments by explicitly controlling for particular deflationary alternative interpretations in their studies, and by arguing that the general strategy behind these arguments fails on theoretical grounds, in that it is inevitably *post hoc* and offers few specific predictions (Scott & Baillargeon, 2014; Scott, 2014). Although the issue remains controversial, it has been nativists, and not their critics,

who have reliably generated new, surprising results in this domain, pushing the boundaries of both how early we thought infants would demonstrate false belief understanding (Kovács et al., 2010; Southgate & Vernetti, 2014) and the sophisticated ways in which they are able to apply this understanding (F. Buttelmann, Suhrke, & Buttelmann, 2015; Moll, Kane, & McGowan, 2015). Proponents of infant false-belief understanding, it would seem, are driving the empirical agenda.

Empiricists might respond to this claim by pointing out the broad success of domain-general models of infant learning (Ruffman, 2014). There is, for instance a growing literature showing that even very young infants are in fact highly sensitive to the statistical regularities in their environments, and that they draw on these regularities when forming expectations about future events (Dewar & Xu, 2010; Kidd, Piantadosi, & Aslin, 2012; Kirkham, Slemmer, & Johnson, 2002; Paulus et al., 2011; Stahl & Feigenson, 2015; Xu & Garcia, 2008). For some critics of theory-of-mind nativism, this growing trend in developmental psychology provides a general theoretical justification for skepticism about infant false-belief claims: if infants really are "intuitive statisticians" (Xu & Garcia, 2008), and there are good reasons for believing that learning in general relies upon tracking observable correlations and inferring hidden causal variables in the environment, then we seem to have good reason to expect that infants rely upon this form of learning during false-belief tasks as well. Thus, while nativists may claim that their interpretations of the new infant data are more predictively fruitful in the restricted domain of theory-of-mind, deflationists may counter that their own interpretation is more coherent with a much broader research program within developmental psychology. Ultimately, this clash of

Quinian virtues seems to result in an impasse, with both sides of the debate firmly entrenched in their respective positions.

However, as we have seen, there is no *prima facie* incompatibility between nativism and domain-general learning. Moreover, no nativist account of mindreading would propose that the mindreading system is innately endowed with *all* there is to know about the world. It is widely acknowledged across the mindreading literature that in order for the mindreading system to function properly, it has to be able to access the subject's general knowledge (Heal, 1996; Nichols & Stich, 2003); mindreading nativism is no different. Even a mindreading system with a rich, innately specified set of general principles for generating hypotheses about the causes of behavior must still be sensitive to regularities about the local environment, or else it would never be capable of giving specific predictions that would actually be relevant to the subject's interests. For instance, in order to predict whether a person will choose to take public transit or her car to go to the zoo on a Sunday afternoon, we need to know something about the woman's preferences and goals; but we also need to know about where the zoo is, whether the public transit system is reliable, the cost of parking, and so on. If it turned out that this knowledge was acquired via domain-general statistical learning, this fact would be entirely consistent with mindreading nativism.

Nativists could also agree with empiricists that we use these statistical learning tools to track regularities in behavior, and to facilitate behavioral predictions. For if domain-general learning mechanisms can track other sorts of environmental realities, there is no reason why they should not also track the regularities that

underlie common human behaviors as well. But while empiricists propose that we parse such regularities as disparate correlations between bundles of low-level perceptual features, the nativist would argue that we instead carve up these regularities in terms of their underlying mental causes. This would mean that instead of acting on hypotheses framed in purely behavioristic predicates (e.g. "agents tend to search for objects that they have repeatedly made contact with in the location where that last contact took place"), as has been argued by deflationists (e.g. Perner, 2010), infants would act on hypotheses framed in mentalistic predicates (e.g. "agents tend to search for objects they desire where they believe those objects to be"). Thus, the difference between empiricists and nativists need not turn on the fact that infants sometimes employ domain-general statistical learning strategies when reasoning about behavior. Rather, the real difference concerns the contents of the hypotheses that infants use to interpret behavioral inputs. Nativists hypothesize that infants learn about behavior by generating and testing mentalistic hypotheses that invoke primitive mental-state concepts and core knowledge about the mind. Empiricists argue that infants learn about behavior by generating and testing hypotheses that invoke non-mentalistic concepts or domain-specific psychological knowledge.

At this point, an empiricist might appeal to considerations of parsimony: her proposal only invokes those representational primitives that we use for any other form of perceptually based reasoning; the nativist, in contrast, must posit additional, mentalistic primitives. Intuitively, the former is more parsimonious. However, appealing to parsimony at this point would be premature: first, it must be established that the two theories under consideration are both empirically adequate, and actually

succeed in explaining the phenomena in question. If one theory cannot account for the relevant *explananda,* while an alternative theory can, then the fact that the former is simpler is irrelevant. In the remainder of this section, I will argue that empiricist accounts of infant false-belief task performance are in fact empirically inadequate, because these accounts do not propose adequate constraints on the learning process.

Before determining whether the empiricist proposal is empirically adequate, let us allow that the empiricist learner can generate new hypotheses about the causes of behavior through the use of causal Bayesian networks. This means that the learner's latent hypothesis space will contain hypotheses about every possible causal pattern (including those with hidden variables) that can be modeled using such tools. This, in principle, should give the child the ability to generate causal hypotheses capable of predicting simple behaviors like the ones that they observe in infant false-belief tasks. Now suppose that such a learner observes a series of behaviors over the course of several familiarization trials in such a task. On the basis of these observations, which hypotheses will the child generate in order to predict the agent's future behaviors?

The problem with this proposal is that there is an immense range of possible (non-mentalistic) causal hypotheses that would be consistent with the child's observations she could potentially generate. For instance, the child could entertain the possibility that the weather is a relevant causal variable for predicting behavior. This wouldn't be an outlandish possibility: the weather is readily observable, and it affects how we behave all the time (e.g. the clothes we wear, whether we go outside, what we do outside, etc.). The child may thus assume that the weather modulates some

hidden causal variables that lead to certain patterns of observable behavior. Thus, the child could formulate numerous hypotheses about the causes of a given behavior that include the weather as a variable. Upon witnessing a familiarization event, the child might therefore entertain the hypothesis that, "agents tend to search for objects that they have repeatedly made contact with in the location where the last contact took place *when it is sunny outside*." Or they might instead focus on some other observable feature of the situation: the color of the experimenter's shoes, the smell of the testing space, whether they have their dolly, and so on. All of these factors might seem like potentially relevant variables, and could be used in the construction of hypotheses about why certain events have occurred. Sometimes, these hypotheses would happen to generate correct predictions. But more often, they would be utterly spurious.

In effect, the empiricist proposal falls prey to a specific version of a familiar problem for empiricist reasoning: Nelson Goodman's new riddle of induction (Goodman, 1955). Goodman's famous insight was that experience could support any number of accidental hypotheses based on spurious predicates. Most famously, Goodman pointed out that the same observations that would support the generalization, "All emeralds are green," could also support the generalization, "All emeralds are *grue*," where *grue* stands for the property of being green until some (distant) time *t*, and blue thereafter. The new riddle of induction asks us how it is that we successfully arrive at solid hypotheses such as, "all emeralds are green," and avoid spurious ones that employ grue-like predicates.

The challenge for the empiricist, likewise, is to explain how it is that children reliably predict behavior in these scenarios, when there is nothing in their model that

would prevent children from generating grue-like hypotheses. How is it that children somehow alight upon precisely the right variables that enable them to make successful predictions, while ignoring all the spurious ones? To answer this challenge, the empiricist needs to posit additional constraints on how children generate hypotheses about behavior, beyond the existence of a domain-general capacity to represent causal patterns. Otherwise, empiricists cannot claim to have provided a complete explanation of how infants are able to reliably predict agents' behavior.

The nativist can offer a clear answer to this challenge: the child ignores spurious generalizations because she is able to evaluate the importance of certain features of the situation in terms of their *relevance* to the target agent's beliefs and desires. By attending to agents' mental states, a child can avoid considering spurious hypotheses based on irrelevant situational factors, and instead focus on those factors that feature in the agent's decision-making process. The child's theory of mind thus provides her with a basis for determining the salience of different features of the environment as she learns about the social world.

This argument recapitulates the basic argument for nativism that I outlined above, as it applies to mindreading specifically. What nativist proposals offer and empiricists lack is a basis for generating and testing informative (rather than uninformative) hypotheses about the causes of behavior. Without domain-specific constraints on the kinds of hypotheses that a child is likely to form about the causes of behavior, empiricist models have no way of explaining why children do not make wildly inaccurate predictions. Core knowledge systems about minds, on this view,

provide children with an overhypothesis about the prior probabilities of different kinds of hypotheses that they could generate to learn about behavior. Spurious hypotheses that make reference to irrelevant situational factors can be safely ignored, because they have extremely low prior probabilities. This resource is unavailable to the empiricist *ex hypothesi*, since the empiricist model treats hypotheses about the causes of behavior like the causes of any other observable event.

To be fair, the nativist proposal is also incomplete. Much more work needs to be done to discover the nature of the representations and inferential processes that guide infant behavior-prediction. However, both the current state of the evidence and the theoretical considerations that I and other nativists have raised give us good reason to believe that early-emerging representations of beliefs and other mental states will be a part of the explanation of the data.

6. Further challenges

6.1. How social experience affects false-belief task performance

Given these considerations, I adopt nativism about belief reasoning – and about mindreading in general – as a plausible theoretical starting point for investigating the architecture of theory of mind more generally. But despite the promise of the nativist explanation of infants' behavior-prediction abilities, such an account faces two parallel empirical challenges. One points to the inadequacy of current nativist explanations of why younger children still fail traditional, explicit false-belief tasks. The second comes from an alternative approach that proposes a "two-systems" account of children's theory-of-mind development.

The typical nativist explanation for why children below the age of four-and-a-half fail the standard false-belief task appeals to the processing burdens that these tasks place on children's developing executive abilities, which prevent children from expressing their underlying competence in belief reasoning (Baillargeon et al., 2010; Leslie & Polizzi, 1998). But while there is some indication that executive factors do contribute to children's performance on these tasks (Benson & Sabbagh, 2005), a recent meta-analysis of this literature has shown that executive functioning only explains a small portion of the variance in false-belief explicit-task performance (Devine & Hughes, 2014). Even more problematically, there is a wide range of evidence suggesting that various socio-linguistic factors, including having older siblings (Ruffman, Perner, Naito, Parkin, & Clements, 1998), exposure to mental-state discourse (Symons, 2004), and training (Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003; Slaughter & Gopnik, 1996) all affect when children pass the false-belief task. Most strikingly, late-signing deaf children (born to hearing parents and not exposed to sign language at an early age) exhibit significant delays on the false-belief task (Pyers & Senghas, 2009; Wellman & Peterson, 2013). Appeals to processing demands cannot explain why these factors influence performance on the false-belief task, as they are largely unrelated to executive functioning. Instead, this evidence suggests that experiential factors play an important role in explaining why children fail (and then subsequently pass) the false-belief task.

In Chapter 2, "Pragmatic Development and the False-Belief Task," I propose a novel nativist account for the role of socio-linguistic experience in children's performance on the false-belief task. The core proposal of this account is that

children's failure on this task is not due to a breakdown in processing resources, but rather an error of Gricean inference. That is, children fail this task because they misinterpret the experimenter's query, not because they are overwhelmed by the demands of the task itself. This pragmatic error is related to children's exposure to belief discourse. As I argue in the paper, children's early input for belief-discourse is pragmatically complex, and initially leads children to hypothesize that beliefs are unlikely to be implicated in conversation. As children's experience with belief discourse increases, they gain more experience with its unique pragmatic elements, and revise their priors about the likelihood of beliefs as a topic of conversation. This proposal has the resources to explain the influence of a variety of experiential factors on children's false-task performance, and also makes a number of concrete empirical predictions.[1]

The key thing to note about this account is how it differs from previous nativist proposals. Whereas previous models attribute children's failures on theory of mind tasks to an a-rational flaw in their underdeveloped cognitive resources, the pragmatic development account assumes that the mechanisms supporting younger children's reasoning in the false-belief task are rationally coherent, and that their responses are justified given their experiences. On this view, children's interpretations of the false-belief task scenario reflect statistically driven priors about what people tend to talk about, and about the kinds of conversational interactions they tend to have. Likewise, their shift in performance during the fifth year of life reflects a rational updating procedure in response to a shifting evidential base. The primary

[1] In the Appendix, I include a paper written with Peter Carruthers in which we show how this account can also be extended beyond the false-belief task, to explain children's developing performance on a number of different measures in the theory of mind literature.

advantage of this approach is that it allows for a nativist explanation that predicts environmentally driven variance in performance, whereas previous nativist explanations could only appeal to internal, maturational factors. Thus, the pragmatic development reflects my earlier point that there need be no contradiction between the facts that children are statistical learners, and that they are innate mindreaders.

6.2. The two-systems account

One of the major problems for the empiricist account of infant false-belief implicit-task performance was that it could not provide an account of the hypothesis-construction systems that underlie infants' accurate behavioral predictions. But there is an alternative to the robust mentalistic interpretation that I have defended that does not share this problem: the two-systems account (Apperly & Butterfill, 2009; Apperly, 2011; Butterfill & Apperly, 2013). In terms of the nativist-empiricist debate, this view is something of a hybrid. Its explanation of how children eventually pass the standard false-belief task is traditionally empiricist: children acquire the concept of belief after their fourth birthdays as the result of a protracted period of social learning, executive development, and experience with language. But rather than explaining infants' false-belief task performance as the product of low-level associative procedures, two-systems theorists accept that it is a genuine form of mindreading. However, they deny that these tasks show that infants actually possess a representational theory of mind. Instead, they posit that it is the product of a more limited, *implicit* mindreading system that tracks beliefs in an extensional sense, but does not actually represent them as such. This system, they suggest, develops prior to and persists into adulthood alongside our mature, representational theory of mind.

Instead of the concept of belief, which represents relations between agents and representational contents, two-systems theorists propose that infants pass false-belief tasks because they have the ability to track *registrations*, which consist in polyadic relations between agents, objects, locations, and times. Similarly, they posit that infants also represent goals and perceptual encounters as purely external, spatial relations that hold between agents and their environments. Two-systems theorists also posit a limited set of predictive rules that characterize how goals, perceptual encounters, and registrations combine to produce behavior. Collectively, this set of quasi-mentalistic concepts and rules forms a discrete, automatic mindreading system that guides infants' passive expectations about behavior. However, the system is not integrated with action-planning or language-production systems, and so it is not able to influence children's responses on explicit false-belief task measures.

One key prediction of this account is that while infants and young children might succeed in predicting behavior driven by false beliefs about objects' locations, these abilities should be subject to *signature representational limits*. For while the registration concept can encode information about where and when an agent encountered an object, it cannot encode information about the *way* agents represent objects. In other words, infants should not be able to predict behavior in Frege cases, wherein an agent encounters an entity under two different modes of presentation, but does not recognize that it is the same entity in both cases (e.g. when Lois Lane encounters Superman and Clark Kent, but does not know that Superman is Clark Kent). A number of developmental studies appear to confirm this prediction (Low et al., 2014; Low & Watts, 2013; Rakoczy, 2015; Surtees, Butterfill, & Apperly, 2012),

although their interpretation is a matter of some controversy (Carruthers, 2015c; Christensen & Michael, 2015; Michael & Christensen, 2016; Thompson, 2014).

Another key prediction of this account is that the dissociation between infant and older children's theory-of-mind abilities should be paralleled in adults. Because the implicit mindreading system exists alongside the explicit one in adulthood, we should expect to see cases where the two dissociate. Specifically, when behavior-prediction only requires the use of the implicit mindreading system, we should expect adults to engage in automatic behavior prediction, which ought to manifest itself in anticipatory looking behavior and in reaction time measures. But when behavior-prediction requires the representation of complex, representational states, we should expect prediction to be slower and more effortful. Once again, two-systems theorists have designed a number of experiments that seem to bear out these predictions (Qureshi, Apperly, & Samson, 2010; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; Surtees et al., 2012; Surtees, Samson, & Apperly, 2016); and again, the interpretation of these findings is disputed (Carruthers, 2015a; Christensen & Michael, 2015; Elekes, Varga, & Király, 2016).

Unlike standard empiricist explanations, this account makes real predictions, and it offers a specific account of the representational basis of the construction of hypotheses about behavior. It has also generated a broad array of empirical results to support it. Empirically speaking, it is on a much firmer foundation than the standard empiricist line. But this foundation is only as stable as the theory of cognitive architecture that motivates its core predictions. One of the things that I do in Chapter 3, "Spontaneous mindreading: A problem for the two-systems account," is analyze

the underlying architectural claims of the two-systems account, and expose the assumptions that drive their empirical predictions.

A key thing to note about the two-systems proposal is that the implicit system is modular in the strong sense – i.e. it is informationally encapsulated (Apperly, 2011). The signature representational limits on the implicit system – specifically, its inability to represent beliefs as such – are a direct consequence of its encapsulation. To represent beliefs, the two-systems theorists argue, requires access to long-term memory and working memory, since BELIEF is a flexible representation that should be able to embed any content whatsoever, and should be able to figure in "promiscuous" inferential relations. In other words, reasoning with BELIEF is something that only a central, "isotropic" system could do (cf. Fodor, 1983). In contrast, reasoning with REGISTRATION only requires the information and specialized inputs that can be processed by the implicit mindreading module. This makes the implicit system fast and efficient, but it also makes it inflexible, which is what explains its signature representational limits. Conversely, the explicit mindreading system, which is able to represent BELIEF, can only do so at the cost of significant processing resources, since these are required to apply such a sophisticated concept.

The central assumption of this argument is that there is a sharp trade-off between a system's capacity for speed and efficiency, and its capacity for representational flexibility. As I point out in Chapter 3, however, this claim is not true of many cognitive systems, including pre-attentional visual processes (once a paradigm case of informational encapsulation). It is thus not obvious that a fast and

efficient system should be unable to represent concepts like BELIEF, even if this does in fact require access to long-term memory. I then go on to discuss evidence showing that the implicit mindreading system is unencapsulated from goals and long-term memory, while the explicit system can be fast and efficient. I conclude that the architectural assumptions driving the two-systems account are false, and that we should abandon it as a framework for interpreting both developmental and adult mindreading data.

6.3. The bigger picture

Explaining the development of the concept of belief is all well and good. However, we also need to be able to translate what we learn from these arguments into a broader understanding of how mindreading works. We need to be able to say how experience and innately channeled mindreading systems combine to support mature mindreading abilities. We should also be able to say something about other forms of mental-state attribution, besides belief-desire reasoning. The two-systems view attempted to do this by offering a comprehensive account of both the development and architecture of human mindreading; but as we have seen, this account comes up short. What can we offer in its place?

Two ideas from Chapters 2 and 3 inform my thinking on this question. The first is reflected in the way the pragmatic development account envisions mental-state attribution as a kind of statistically informed Bayesian hypothesis-testing. The account assumes that we use data from the environment to generate prior probabilities for different mentalistic hypotheses, which we then use to predict and interpret behavior. This also reflects the more general point that nativism is perfectly

compatible with domain-general forms of learning. But most importantly, this kind of model has ample resources to explain experience-dependent variance in mindreading performance, which a broader account should be able to do.

The second idea comes from the observation in Chapter 3 that the extent to which we track and update representations of other agents' mental states depends in part upon our motivations. This is a deep point: generating new hypotheses about the mental states can involve varying amounts of attention and other cognitive resources. Depending on the context and the identity of the target, mindreaders may sometimes be better served by diverting those cognitive resources elsewhere. Mature mindreading performance seems to reflect this fact by selectively updating agents' mental states when doing so is relevant to the mindreader's goals.

This kind of goal-sensitive, statistically informed, probabilistic approach to mental-state inference is on full display in the mindreading architecture that I put forward in Chapter 4, "Character and theory of mind: An integrative framework." The stated goal of this paper is to better understand a form of mindreading that is often ignored in the theory-of-mind literature, namely, character-trait attribution. But I also use the opportunity to develop a broader account of how different forms of mental-state representations contribute to the prediction and interpretation of action. I do this by invoking a hierarchical predictive processing architecture, a model of neural information processing that has become highly influential in a number of areas of cognitive science, including theory of mind (Clark, 2015; de Bruin & Strijbos, 2015; Hohwy & Palmer, 2014; Hohwy, 2013; Koster-Hale & Saxe, 2013; Rao & Ballard, 1999; Spratling, 2016). Importantly, this kind of architecture is built around precisely

the kind of Bayesian hypothesis testing that I have argued governs mental-state inference. It also has the resources to explain how motivational factors can modulate the extent to which cognitive resources are devoted to particular representational tasks. In other words, hierarchical predictive coding provides an ideal framework for modeling the architecture and development of mindreading.

Of course, the model I describe in Chapter 4 is of mature mentalizing; it makes no explicit claims about how any of these abilities develop, nor does it offer any particular support for nativism. But it does cohere with the overarching Bayesian approach that I advocate throughout my dissertation. It does not propose that low-level, perceptual associations or a fixed set of quasi-mentalistic behavior rules predict behavior. The picture of mindreading it offers instead is abstract, highly structured, and able to flexibly respond to evidence. Moreover, the particular structure for mental-state inference that it proposes - a predictive hierarchy that moves from temporally stable states to transient ones - reflects the hierarchical structure of intentional action itself. Thus, what it offers for nativists is a picture of what our innate mindreading capacities become, a model for what a mature, rationalistic theory of mind might look like. With such a picture in mind, nativists can start to think further about how the core hierarchical structure of action prediction and interpretation develops over time.

7. Conclusion

The philosophical roots of the theory of mind debate can be traced to the problem of other minds, but my specific approach to it has been most deeply influenced by the debate between nativists and empiricists. I've suggested that if we think of learning on the model of Bayesian hypothesis testing, then we can think of the nativist's proposed innate learning structures as providing constraints on hypothesis-generation that make learning possible. Without such innate constraints in the domain of mindreading, I've argued, we simply cannot make sense of young children's precocious mental-state attribution abilities, especially on "implicit" false-belief tasks. In other words, innate, domain-specific constraints on mental-state attribution are what make children's learning about other minds possible.

Given this starting point, the nativist about theory of mind must address a key challenge: the gap between infants' "implicit" false-belief understanding and older children's explicit false-belief competence. Chapter 1 offers a novel pragmatic explanation for why this gap arises (which I expand upon in the Appendix), while Chapter 2 challenges the underlying architectural commitments of an alternative view, the two-systems account. Chapter 3 takes a broader view, and puts forward a proposal about the structure of mature mental-state inference (while at the same time explaining the role of character-trait attributions in mindreading).

Chapter 2: Pragmatic development and the false-belief task[2]

1. *Introduction*

Since it became a topic of empirical research, the study of children's theory of mind – their understanding of the underlying psychological basis of behavior – has been dominated by the discovery that younger children systematically fail false-belief tasks, and start to succeed sometime after their fourth birthdays (Wellman et al., 2001; Wimmer & Perner, 1983). The debate regarding the interpretation of this discovery has divided philosophers and psychologists along nativist and empiricist lines. Empiricists have claimed that the shift in performance on false-belief tasks around children's fourth year signals their acquisition of a genuinely meta-representational concept of *belief* (Gopnik & Wellman, 1992; Perner, 1991). Nativists argued that younger children's failures reflected a performance error related to children's underdeveloped executive and attentional resources and the processing demands inherent to the task, rather than a fundamental lack of competence with the concept of *belief* (Fodor, 1992; Leslie, Friedman, & German, 2004).

In the two decades after the false belief task was first introduced as a measure of theory of mind development, both empiricist and nativist camps remained firmly entrenched (see, for example, Scholl & Leslie's (2001) response to Wellman et al. (2001)). More recently, new methods for studying false belief understanding in preverbal infants appear to have vindicated the nativist position (Barrett et al., 2013;

[2] This chapter was originally published as Westra (2016). It has been reprinted here with permission from Springer, copyright license # 4067640601646.

D. Buttelmann et al., 2009, 2014; Kovács et al., 2010; Onishi & Baillargeon, 2005; Senju et al., 2011; Southgate & Vernetti, 2014). These studies seem to show that while younger children do systematically fail false-belief tasks that attempt to elicit explicit, communicative responses, infants as young as 6 months of age understand false beliefs in tasks where success is measured by their spontaneous reactions to behavior, either with anticipatory looking, violation of expectation, active helping, or EEG paradigms.

Interpreting these findings has created a great deal of controversy. Many authors have argued that such implicit measures do not actually demonstrate genuine meta-representational abilities, and offered a variety of alternative interpretations that preserve the empiricist narrative (Butterfill & Apperly, 2013; Gallagher & Povinelli, 2012; Heyes, 2014a; Perner, 2010). In response, nativists have produced a steady stream of new empirical results aimed at refuting these deflationary hypotheses (D. Buttelmann et al., 2014; Moll et al., 2015; Scott, Richman, & Baillargeon, 2015; Senju et al., 2011). Others have defended a rich, mentalistic interpretation of the new infancy data on theoretical grounds, pointing out that the *post hoc* nature of many of these deflationary proposals counts against their credibility, while other proposals seem to be ill-equipped to explain the sheer range and flexibility of infants' socio-cognitive abilities (Baillargeon et al., 2010; Carruthers, 2013; Christensen & Michael, 2015; Scott & Baillargeon, 2014; Scott, 2014; Thompson, 2014). Thus, in spite of this controversy, theory-of-mind nativism continues to enjoy substantial evidential support, and is currently driving a highly productive research program. There is, in

short, good reason to think that children possess the concept of belief well before they pass the false belief task.

However, even if we accept this conclusion (as I will in this paper), nativism about theory of mind development still faces a significant challenge when it comes to explaining why younger children systematically fail the standard false belief task (hereafter, FBT). The standard nativist line is that these tasks impose severe demands on young children's still-developing executive resources, which causes them to fail. But this account is ill-equipped to explain why certain forms of social experience and training affect when children succeed on the FBT, as it is not clear that the findings in question could be explained solely in terms of a child's developing executive abilities.

The goal of the current paper is to show how these findings fit into a revised nativist framework. Typically, data showing a role for social experience in theory-of-mind development are cited in support of empiricist accounts. However, there is no inconsistency between nativism and a role for social learning. All contemporary nativist approaches to the mind are meant as explanations for *how* individual learning takes place; they do not deny that individuals ever learn at all, or that innate knowledge is ever enriched by experience (*pace* Fodor (1975)). But it is incumbent upon the nativist to explain how various types of experience lead to individual differences in theory of mind development.

This is the challenge that I take up in this paper. Given the strong evidence in its favor, and its growing acceptance in the field, I will be taking the claim that young children can represent beliefs as a point of departure. Those who are agnostic or

skeptical of this claim are invited to view the account I'll be laying out as a way of filling in the following conditional: *if* infants *could* represent beliefs, how would we go about explaining the influence of social experience on FBT performance? Providing the best answer to this conditional question should be important even for those who ultimately reject its antecedent.

My proposal, which I'll call the *pragmatic development account*, is that while young children are capable of representing beliefs early on in development, they are not yet very good at understanding when facts about belief are relevant to conversation. In spite of the fact that they constantly attribute beliefs, desires, goals and intentions to other agents, understanding when these pre-linguistic concepts are implicated in conversation is not just a matter of acquiring the right vocabulary. Young children do not initially expect people's beliefs to be a topic for conversation. They have to learn this through experience with the pragmatics of belief discourse – that is, during social interactions in which facts about beliefs are implicated in conversation. With this experience, children learn to adjust their prior expectations about the relevance of doxastic facts when interpreting particular speech acts.

As a result, different levels of exposure to belief discourse can affect how children interpret questions like the ones they must answer in FBTs. When they lack the requisite experience, children are prone to misinterpret the crucial false belief query as a kind of indirect speech act that is not about the beliefs at all. But as they gain more experience with belief discourse, children start to recognize the true purpose of the experimenter's question, and respond accordingly. In other words,

younger children fail the FBT not because they lack the concept of *belief* or because the tasks are too executively demanding, but due to a mistaken Gricean inference.

The pragmatic development account is not wholly new. Siegal and Beattie (1991) proposed a Gricean account of younger children's systematic failure on FBTs. They argued that three-year-olds are typically too inexperienced to pick up on experimenters' conversational implicatures during the FBT; as a result, they fail to grasp the relevance of mentalistic factors to the experimenters' questions, opting instead for a more familiar, world-oriented interpretation. Thus, when children hear "Where will Sally look for her marble?" they interpret it as, "Where will Sally have to look for the marble *in order to find it*?" rather than "Where will Sally look for her marble *first*?" Siegal and Beattie supported this interpretation by showing that three-year-olds tended to pass a modified version of the FBT in which they were asked the latter question, even though they would still fail when asked the former. Later, Surian and Leslie (1999) both replicated Siegal and Beattie's findings and expanded upon them by showing that a similar manipulation failed to improve the performance of a control group of individuals with autism spectrum disorder (a population widely believed to suffer from a chronic theory of mind deficit). In support of a similar hypothesis, Hansen (2010) argued that children in the FBT might interpret the experimenter's query as a question about the state of the world, rather than a question about the agent's mental states. To this end, he showed that young children perform much better on FBTs in which the experimenter makes it clear the he is not asking about the state of the world: "You and I both *know* where Sally's marble is, but where does Sally *think* it is?" Pursuing a different version of the pragmatic development

strategy, Helming and colleagues have proposed that it is children's propensity to be helpful which leads them to misinterpret the experimenter query during the FBT (Helming, Strickland, & Jacob, 2014). The current proposal builds upon these earlier pragmatic accounts in that it specifically engages with the social learning that goes into passing the FBT, rather than simply focusing on the on-line pragmatic demands of the task itself. It also makes a number of novel recommendations about how to tease apart the respective contributions of children's executive abilities and their pragmatic understanding of the task.

In the next section, I describe current nativist explanations of children's failure on the FBT, and then present findings that seem *prima facie* inconsistent with these explanations. In the third section, I lay the groundwork for the pragmatic development account, and present several arguments for why belief discourse poses pragmatic challenges for the novice speaker. In section 4, I present the core elements of the pragmatic development account, and show how it is able to explain children's performance on a wide range of FBTs. In section 5, I show how the account is able to accommodate a particularly challenging set of findings from Call and Tomasello (1999). In section 6, I end by making several predictions that would distinguish my own view from other nativist accounts.

2. A challenge for existing nativist accounts

Many of the prominent nativist accounts of theory of mind development have focused on the processing load that the FBT places on younger children's developing executive functioning. Baillargeon and her colleagues' *response account*, for instance, posits that younger children are unable to cope with the demands of

simultaneously attributing a false belief, selecting a response to the experimenter's question, and inhibiting a prepotent tendency to answer the experimenter's question with her own knowledge, perhaps due to still immature connections between mindreading and executive regions of the brain (Baillargeon et al., 2010). Along similar lines, Leslie and colleagues have argued that success on FBTs is modulated by the development of a domain general *selection processor* responsible for inhibiting the mindreading system's tendency to attribute the subject's own beliefs to others by default (Leslie, German, & Polizzi, 2005; Leslie & Polizzi, 1998). Carruthers (2013) also holds a "processing load" view, but emphasizes that all three components of FBTs – attributing a false belief, interpreting the experimenter's question, and generating a response that will communicate the appropriate information to the experimenter – involve mindreading (see also Sperber & Wilson, 2002). According to this *triple mindreading account*, executing each of these tasks simultaneously places heavy demands on both processing resources internal to the mindreading system and general executive resources, both of which may be insufficiently developed in younger children, which explains why younger children fail the FBT while still possessing the concept of *belief*.

In line with these "processing load" accounts, a number of studies have found that performance on various executive tasks predicts earlier success on the FBT (Benson & Sabbagh, 2005; Carlson, Moses, & Breton, 2002). However, the overall correlation between these two constructs is in fact quite weak. According to a recent meta-analysis by Devine and Hughes, the correlation between executive functioning tasks and performance on FBTs is only .22 after controlling for age and verbal ability,

with differences in executive functioning accounting for only 8% of the variance in performance on FBTs (Devine & Hughes, 2014). Moreover, this correlation was consistent across diverse measures of executive functioning. This suggests that no single component of executive functioning accounts for its relation to FBT performance. Thus, although it does make a small contribution to children's performance on the FBT, it seems unlikely that executive functioning holds the key to explaining why most children fail the task until after their fourth birthday.

Moreover, any account that appeals solely to the maturation of children's executive abilities as an explanation of how they come to pass the FBT is ultimately underequipped when it comes to explaining the various experience-related factors that influence FBT performance. For instance, it's been shown that the extent to which a child's mother talks about mental states predicts how early that child will succeed on FBTs (Ruffman, Slade, & Crowe, 2002; Symons, Fossum, & Collins, 2006; Symons, 2004). Beyond maternal interactions, children with older siblings also appear to have an advantage on the FBT (Perner, Ruffman, & Leekam, 1994; Ruffman et al., 1998). Further, interventions that train children on various aspects of mental state discourse have tended to improve children's performance on FBTs (Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003; Slaughter & Gopnik, 1996; Wellman, 2012).

Exposure to language in general also has dramatic effects on when children are able to pass the FBT. Deaf children born to hearing parents who are exposed to sign-language late in life are significantly delayed on explicit false-belief tasks when compared to both hearing children and deaf children born to deaf parents (whose FBT performance is comparable to that of hearing children) (Peterson, Wellman, & Liu,

2005; Wellman, Fuxi, & Peterson, 2011). Notably, this delay is not the result of any sort of congenital neurological abnormality (as is the case with children on the autism spectrum) but is instead due to purely environmental factors. Nevertheless, late-signing deaf children still reliably display the same developmental progression through various types of theory-of-mind problems as typically developing children (e.g. succeeding on problems involving diverse desires before problems involving false beliefs; see Section 4.1). However, late-signing deaf children are able to succeed earlier on FBTs after they are exposed to theory-of-mind-based interventions using "thought bubbles" that draw attention to individuals' beliefs (Wellman & Peterson, 2013).

Some of the most striking evidence for the importance of experiential factors in theory-of-mind development comes from a natural experiment that took place in Nicaragua during the last few decades of the 20th century. In 1977, an expanded elementary school for special-needs children was opened in the city of Managua. Here, for the first time, deaf children in Nicaragua came into extended contact with one another. Although their education was conducted in Spanish, amongst themselves the students began to develop their own novel system of gestural communication, an amalgam of the children's various idiosyncratic home-sign gestures. This system of gestural communication was expanded as older students passed it on to new ones, and rapidly developed into a full-fledged sign language known today as Nicaraguan Sign Language, or NSL (Senghas, Kita, & Ozyürek, 2004).

Importantly, the version of NSL acquired by its earliest speakers was less complex than the one acquired by later speakers, and completely lacking in mental

state vocabulary (Pyers & Senghas, 2009). In a study with adult speakers of NSL that compared the performance of earlier "first cohort" and later "second cohort" speakers of NSL, Pyers and Senghas found that first cohort speakers systematically failed a non-verbal elicited-response version of the FBT, while second cohort speakers were generally successful. In a follow-up several years later, the performance of the first cohort speakers on the FBT had significantly improved. Pyers and Senghas attributed this improvement to an intermingling between first and second cohort speakers of NSL, leading the first cohort speakers to acquire a greater facility with mental state discourse. Note that one could not plausibly attribute the change in the first cohort speakers' performance on the FBT to a development in executive abilities (as the nativist might for the parallel change in performance of 3-4 year olds), as these subjects were *adults* at the time of the first test, and likely possessed fully mature executive resources. Indeed, both the difference between first and second cohort NSL speakers and the change in first cohort speakers' performance appear to be the result of social experiences specifically related to their acquisition of mental-state vocabulary.

Explanations of FBT performance that appeal solely to the on-line demands that the task places on executive resources do not tell us much about why these kinds of experiences affect when an individual ultimately overcomes those demands. Even if important maturational changes to children's executive resources do occur between the ages of four and five, and individual differences in executive functioning do correlate somewhat with individual differences on the FBT, it's not obvious how these internal cognitive developments could explain why an individual's social

experiences also seem to matter for her performance on the FBT, particularly when that individual does not pass the FBT until long after her fifth birthday (as is the case with language-deprived children). This suggests that a child's social environment makes an independent contribution to her performance on the FBT.

At this point, one might suggest that language might be the crucial factor in passing the FBT. Indeed, various authors have proposed a crucial role for various aspects of language in theory-of-mind development, including complementation syntax (de Villiers & Pyers, 2002), mental-state vocabulary (Montgomery, 2005), and the social experience that comes with linguistic interactions (Dunn & Brophy, 2005; P. L. Harris, de Rosnay, & Pons, 2005; Tomasello & Rakoczy, 2003); however, in a recent meta-analysis of the theory-of-mind and language literature, Milligan, Astington, and Dack (2007) were unable to identify a special role for any single aspect of language independent of general language ability. Moreover, after controlling for age, they determined that linguistic factors accounted for roughly 10% of the variance in theory of mind abilities. Thus, language, like executive functioning, makes only a small (albeit statistically significant) contribution to performance on the FBT.

However, one limitation of this meta-analysis was that it did not assess how pragmatic learning affects children's performance on the FBT. In the next section, I explore the pragmatic factors that accompany both the FBT, and belief discourse in general.

3. The pragmatic challenges of belief discourse

To begin to make sense of all the above-mentioned findings within a nativist framework, we must first be clear about the basic problem that contemporary versions of theory of mind nativism are meant to solve, namely, explaining how even very young children's spontaneous expectations about behavior seem to be sensitive to the mental states of others. Nativists posit that they are able to do this because they possess innately channeled inference mechanisms that take observable behaviors as input and generate mental state attributions as output. But this account only explains how young children come to possess mental state *concepts*. Learning to apply these concepts in an adult-like manner in *linguistic interactions* is another story. A novice speaker of a language, even one who is able to represent the mental states of others, may nevertheless demonstrate non-adult-like performance on tasks that require her to interpret other speakers' utterances as being *about* mental states. After all, the nativist's hypothesis is about where our conceptual understanding of mental states comes from, not how we learn to talk about them. The nativist about mindreading is silent when it comes to explaining how we learn to participate in mental-state discourse – which, it turns out, is surprisingly tricky for the novice speaker. In particular, it appears that younger children do not expect *beliefs* to be a likely topic of conversation.

3.1. References to beliefs in the explanation and description of behavior

To see why doxastic facts pose a particular difficulty for the novice speaker, consider first the asymmetrical roles that beliefs and desires play in ordinary folk-psychological explanation (Rakoczy, Warneken, & Tomasello, 2007; Steglich-

Petersen & Michael, 2015). Suppose, for instance, that we observe Sally walk over to the cookie jar and open the lid. When asked why Sally opened the lid to the cookie jar, a natural and perfectly informative response would be, "Because she wanted a cookie." Note that this response makes no mention of Sally's beliefs – just her desires. Now, consider an alternative response: "Because she wanted a cookie, and she believed that there would be cookies inside the jar when she opened it." This explanation, while accurate, is a bit odd. To mention Sally's belief in this context seems to provide too much information, a violation of Grice's Maxim of Quantity (Grice, 1991). Sally's belief about the cookie jar is so obvious that it is simply not worth mentioning. This is because when we give explanations of this type, we tend to presuppose that facts about Sally's beliefs are a part of the conversational common ground. Even when this is not in fact the case, and the listener actually does not take facts about Sally's beliefs to be in the common ground, the speaker's act of only referring to Sally's desires is itself evidence that *some* fact about Sally's beliefs has been presupposed. It is then incumbent on the listener to supply that fact herself in order to render the explanation coherent.[3] Thus, overt reference to beliefs is notably absent from even this very simple instance of a folk psychological explanation; in its place, we find a subtle practice that relies upon presupposition and pragmatic inference.

[3] This is a weak form of *presupposition accommodation*, which takes place whenever speakers dynamically update the set of propositions that are taken to be a part of the common ground in response to changes in the conversational context. Thus, for example, a felicitous utterance of "It was Jon who broke the doorknob" presupposes that the doorknob has been broken, and this leads the listener to infer that "the doorknob has been broken" is now a part of the common ground (Stalnaker, 1998).

Our descriptions of behavior also seem to frequently omit reference to beliefs. In an elegant series of experiments, Papafragou et al. (2007) presented both adults and children with short scenes, which the subjects were then asked to describe. In their control conditions, they found that both adults and children tended to make very few references to the actors' beliefs when describing the scenes, opting instead to refer to agents' goals, or simply to their overt physical behaviors. However, the experimenters hypothesized that both children and adults would be more likely to describe a scene in terms of actors' beliefs when they are provided with additional cues that make doxastic factors more salient. Specifically, they predicted that the presence of syntactic cues from sentences with clausal complement structure (e.g. "Sally believes THAT the marble is in the box,") or situational cues in which a character acts on a false belief would prompt subjects to use more belief words.

To test these predictions, the authors presented both adults and children between the ages of three and five with silent scenarios showing actors engaged in various activities. Some of these scenarios showed actors performing simple actions, while others showed the actors acting on false beliefs (e.g. absent-mindedly drinking from a flower vase that had been placed where their water glass was while they were not looking). In some cases, these scenes were accompanied with nonsense sentences containing either a clausal complement structure introduced by 'that' (e.g. "Vanissa LODS that she ziptorks the siltap"), a transitive structure with a direct object ("Vanissa VAMS the torp"), or an intransitive structure ("Vanissa TROMS"). Across their experiments, they found that both the false-belief scenario and the clausal complement cue substantially increased both adults' and children's references to

beliefs when describing what they saw. This effect was strongest when both cues were co-occurring; when such cues were absent, they tended to describe the scene using non-doxastic vocabulary.

These results show two things: first, that we do not spontaneously refer to beliefs in our behavioral descriptions; second, talk of beliefs seems more likely when some feature of the situation has raised the saliency of belief facts. Thus, in description, as with explanation, doxastic facts are not often mentioned under ordinary circumstances. Yet representations of belief facts still appear to be available, as overt references to them can be prompted by the presence of a syntactic cue. The fact that false-belief scenarios do prompt references to beliefs is also telling, because it suggests that it is only in somewhat unusual circumstances that it becomes important for speakers to draw attention to beliefs. This suggests that while we do represent the beliefs of others, it is only in special circumstances that these representations get overtly mentioned in conversation.

If this asymmetry in the role of beliefs in the explanation and description of behavior were in fact pervasive in the novice speaker's linguistic input, then we would expect a corresponding asymmetry in the frequency of overt references made to beliefs and desires in child-directed speech. There is some indication that this is in fact the case: according to the Child Language Data Exchange System (CHILDES) database, by age 4, children have heard the verb 'think' an average of 611, 220 times, and 'want' 1.3 million times (MacWhinney, 2014).[4] We see something similar in a study conducted by Tamoepeau and Ruffman, in which mothers were made to tell a

[4] Corpus analyses are due to Kaitlyn Harrigan and Aaron White (Department of Linguistics, University of Maryland, College Park).

story to their children from a book containing only images: references to desires were roughly twice as frequent as references to beliefs (Taumoepeau & Ruffman, 2006). These findings provide support for the claim that we frequently omit references to beliefs in our explanations and descriptions of behavior. They also highlight a more basic fact, namely that belief discourse input is relatively sparse for a novice speaker, at least when compared to desire input.[5]

In both our explanations and our descriptions of behavior, then, facts about belief are often left implicit. For adults, this pragmatic dimension of belief discourse is barely noticeable, and engaging in these discursive practices is positively effortless. But for a child – even one who possesses the concept of *belief* – this might make belief discourse rather difficult. Not only must the child be able to grasp the role of beliefs in generating behavior, but she must also know that common knowledge of these facts is often being presupposed during conversation. But until she has learned this, she will only notice that talk of beliefs is comparatively rare. For the child, it will seem as though beliefs are not the sort of thing that people are often interested in talking about.

3.2. The pragmatics of 'thinks'

Another factor adding to difficulties associated with belief discourse is that the verb that we most often use to express the belief concept, 'think,' is generally not used to attribute beliefs. Often, 'think' is used in indirect speech acts as a way of proffering a

[5] There may be other reasons for the prevalence of desire-talk in child-directed speech when compared to belief-talk. For instance, it may be that caregivers query children about their desires much more often than their beliefs because caregivers are more interested in satisfying children's needs than in hearing about what they think.

complement clause that the speaker takes to be true (Simons, 2007). To illustrate, consider the following exchange:

Agnes: When does the game start?

Roberta: I think that it starts around 7pm.

Interpreted literally, Roberta has responded to Agnes' question by self-attributing a belief about the game. But this interpretation would be bizarre: facts about Roberta's mental states are orthogonal to the question under discussion, and Roberta's referring to them would seem to violate the Maxim of Quantity by bringing up irrelevant information. Of course, we do not interpret Roberta's utterance in this manner because it is clear that the primary illocutionary act being performed is not, in fact, about Roberta's mental states, but rather about the game itself.[6] In the exchange above, Roberta is using 'think' as a way of indirectly endorsing the truth of the complement clause, namely, that the game starts at 7pm. Used in this "parenthetical" manner (J. Hooper, 1975; Simons, 2007), sentences of the form "S thinks that P" become pragmatically enriched so that they imply that the speaker takes P to be true; in contrast, literal, attributive uses of "S thinks that P" are neutral with respect to the truth of P.[7]

[6] Similarly, consider how the primary illocutionary act behind the familiar "Could you pass the salt?" is a request for salt, not question about someone's salt-passing abilities (Searle, 1975).

[7] Third-personal instances of "S thinks that P" are also often indirect. For instance, suppose that Roberta were to answer Agnes' query from the dialogue above with "Carlos thinks it starts at 7pm." Once again, the primary illocutionary act in this case is not to draw attention to Carlos' beliefs *per se*, but rather to reply to the speaker's question. Roberta's use of 'think,' in this utterance, serves an evidential function: it provides information about the source of Roberta's reply, and it adds a qualification about the reliability of that source. Thus, we might accurately paraphrase Roberta's utterance here as, "To the best of my knowledge, the game starts at 7pm. I learned this from Carlos" (Simons 2007).

Thus, utterances containing 'think' often require an additional inference about speaker meaning to determine whether it is being used indirectly or attributively, which in turn impacts whether or not the complement clause is being asserted as true. Even worse (from the perspective of the novice speaker), indirect uses of 'think' appear to be far more common than attributive uses: corpus analyses of child-directed speech reveal that the overwhelming majority of adults' uses of 'think' are of the indirect variety; correspondingly, most of younger children's early uses of 'think' tend to be indirect and first-personal in nature, rather than genuine belief ascriptions (L. Bloom, Rispoli, Gartner, & Hafitz, 1989; Diessel & Tomasello, 2001; Shatz, Wellman, & Silber, 1983).The combination of the infrequency with which we overtly refer to beliefs in explanation and description and the pragmatic noisiness of 'think' makes interpreting utterances containing 'think' quite challenging for the novice speaker. It is therefore unsurprising that children below the age of four also sometimes show non-adult-like comprehension of 'think,' and often seem to treat it as equivalent to 'know' (Johnson & Maratsos, 1977; Moore, Bryant, & Furrow, 1989).

Multiple authors have interpreted younger children's difficulties with epistemic verbs as evidence of an underlying conceptual deficit: younger children fail to distinguish the meanings of 'think' and 'know' because they lack the concepts those words express (Perner, Sprung, Zauner, & Haider, 2003; Tardif & Wellman, 2000). However, recent experimental evidence suggests that, contrary to the above interpretation, children do in fact demonstrate an adult-like semantic understanding of 'think', provided that extraneous task demands have been sufficiently reduced.

Second-personal instances of 'think' are often indirect as well. If I ask, "Do you think it's going to rain?" I am effectively asking whether it will rain. Here too, the question under discussion does not concern your mental states, but rather facts about the world.

Specifically, while children do poorly on tasks requiring them to say *what* an individual thinks, they do much better when they are asked to make truth-value judgments about sentences in which 'think' is used attributively (Dudley, Orita, Hacquard, & Lidz, 2015; Hacquard, 2014; S. Lewis, Hacquard, & Lidz, 2012; S. Lewis, 2013).

Pursuing this idea, Lewis, Hacquard, and Lidz (2012) proposed that children's non-adult-like performance on other tasks involving 'think' is the product of pragmatic factors, not a conceptual or semantic deficit. According to this 'pragmatic development hypothesis,' three-year-olds do in fact have the appropriate semantics for 'think' and the corresponding concept of *belief,* but they tend to make incorrect inferences about the intentions behind utterances in which 'think' occurs, treating literal uses of 'think' verbs as indirect by default. This hypothesis predicts that experimental manipulations that make attributive interpretations of utterances containing 'think' more salient should lead to more adult-like performance on comprehension tasks.

To test this prediction, Lewis et al. (2012) presented a sample of four-year-olds with vignettes in which cartoon characters played a game of hide-and-seek. After watching one or more characters hide, participants first interacted with a puppet that would ascribe beliefs to the seeker (e.g. "Dora thinks Swiper is behind the toy box,") and then were asked by the experimenter whether or not what the puppet said was correct. In their first experiment, participants tended to give incorrect truth-value judgments when the puppet accurately ascribed false beliefs to the seeker. However, in their next experiment, a *second* seeker with conflicting beliefs about the location of

the hider was added to the vignette. In this experiment, participants' truth-value judgments about the puppet's belief ascriptions improved across all conditions, revealing an adult-like semantic understanding of the verb 'to think'.

To explain this improvement, the authors suggest that children in the 1-seeker condition failed because they defaulted to an indirect interpretation of the puppet's use of 'think', which led them to infer that the puppet was in fact proffering a false statement. By introducing another conflicting perspective to the scenario, the authors were able to highlight the relevance of the first seeker's beliefs in the child's conversation with the puppet, which led the children to interpret the puppet as using 'think' attributively and give the correct answer. This suggests that the subjects' initial responses were not based on a failure to represent the character's beliefs or an immature understanding of the meaning of 'think', but rather a failure to correctly interpret the speaker meaning behind the original belief ascription made by the puppet.

Notably, standard nativist accounts of children's theory of mind development that stress the development of executive functioning would not have predicted this result. Such an account would have predicted that the addition of the second seeker would have made the task harder, since adding another perspective to the situation would have given the subjects yet another concurrent mindreading task and increased the executive burden of the task. The fact that adding the second seeker did not have this effect is further evidence that demands on executive functioning do not fully explain children's systematic failures on the FBT.

Introducing a telling contrast in order to highlight the attributive interpretation of 'think' has also been demonstrated in three-year-olds. Arguing along similar lines as Lewis et al. (2012), Hansen (2010) showed that younger children's success rates on FBTs surpass chance levels when experimenters ask, "You and I both *know* where Sally's marble is, but where does Sally *think* it is?" This manipulation is particularly effective, since it actually introduces two pieces of contrastive information that serve to highlight the relevance of the subject's doxastic state. First, by drawing attention to the knowledge she shares with the child, the experimenter's query serves to highlight the fact that Sally does not share in this knowledge. Second, the query involves the use of both 'know', which has a factive semantics, and 'think', which does not. Contrasting 'think' and 'know' in this context is an effective way of eliminating the possibility that 'think' is being used indirectly, since such an interpretation (which tends to imply that the complement clause is true) would render the contrast with 'know' uninformative. Thus, both pieces of contrastive information contained in the experimenter's query lead the child to interpret the topic of conversation to be Sally's beliefs rather than reality. Both Hansen (2010) and Lewis et al. (2012) thus provide compelling evidence that children are able to grasp that 'think' expresses a belief attribution, provided that other, non-doxastic interpretations of 'think' are excluded by contextually relevant information.

4. The pragmatic development account

One thing that the studies by Hansen and Lewis et al. tell us is that we should expect younger children to have difficulties on FBTs that ask them what a particular agent *thinks* (e.g. Jacques & Zelazo, 2005; Low & Simpson, 2012): in those tasks, children

are likely defaulting to an indirect interpretation of the verb, rather than an attributive one. Notably, the subset of FBTs that employ the 'thinks'-question includes the majority of unexpected-contents and deceptive-object versions of the FBT (e.g. Perner et al. 1987; Gopnik and Astington 1988).[8] This is because these tasks all try to draw children's attention to their own or another agent's prior expectations about the world, which requires an overt reference to beliefs. Since younger children do not expect beliefs to be a topic of conversation, and are used to 'think' being used in indirect speech acts, they naturally interpret this to be the experimenter's true communicative purpose. They thus interpret the question "what will S think is in the box?" as an indirect question about the contents of the box, and respond accordingly.

4.1. "Where will Sally look for the apple?"

However, many standard FBTs do not ask children what a particular character will *think*, but rather where she will *look* (e.g. Wimmer & Perner, 1983). It might be objected that the above results tell us nothing at all about why younger children fail this kind of task, since the word 'think' is never actually used. But this objection misses the point of the proposal. The frequency of indirect uses of 'think' and the absence of overt references to belief in explanations and descriptions of behavior would, according to this account, lower the probability that belief facts are relevant to interpreting the speech acts of others. This would hold true *regardless* of whether the word 'think' is used in a given utterance. Thus, when an experimenter asks a child,

[8] In an unexpected-contents FBT, children are shown a container that looks like it should contain one thing (e.g. candy), but really contains another (e.g. pencils). They are then asked to say what another ignorant agent will think is in the box, or what they themselves originally thought was in the box before seeing its contents. In a deceptive-object FBT, they are shown an object that looks like one thing (e.g. a rock), but is really another (e.g. a sponge), and then get asked a similar set of questions as in the unexpected-contents task.

"Where will Sally look for her marble?" he wants the child to show that she knows that Sally believes that the marble is in its old location. But, if the child has had little experience with belief discourse, then she is unlikely to judge that this is the speaker's true communicative intention. Because such an interpretation would implicate beliefs as a topic of conversation, its low prior probability would place it at a disadvantage relative to any other, competing interpretations that might be available. Thus, the child would be unlikely to attribute the doxastic interpretation to the experimenter's speech act.

Importantly, I do not claim that the pragmatics of the false-belief task cause children to *lose track* of the agent's beliefs, or that pragmatic factors make this information cognitively inaccessible; this would render my view virtually indistinguishable from a processing-load account. Children, on my account, spontaneously infer and maintain mental representations of the agent's beliefs and goals throughout the false-belief scenario. But although they represent the agent as having certain beliefs, they fail to infer that the *experimenter is interested in* those beliefs, and that this interest is motivating her communicative intention. Thus, while children are perfectly capable of representing false beliefs in this scenario, their inexperience with belief discourse leads them to err in their Gricean reasoning about what the experimenter wants from them.

But while younger children do, from our adult perspective, get the answer to the FBT wrong, it is important to note that from their perspective, their answer is perfectly justified. For given their experience with belief discourse, the actual communicative intention of the experimenter – which is to get children to show that

they know that Sally thinks the marble is in the incorrect location – would seem quite unusual. It is only natural that children should instead attribute to the experimenter a more plausible communicative intention.

But, one might wonder, how *else* would the child interpret the experimenter's query in when she asks where Sally will look for the marble? As proponents of other pragmatic development accounts have pointed out, the standard change-of-location scenario creates a set of conditions in which other interpretations of the experimenter's query would be highly salient to the child. For instance, Siegal and Beattie (1991) suggest that children may interpret the experimenter's query as "Where will Sally *find* the marble?" given that obtaining the ball is Sally's ultimate goal in the FBT scenario, and children treat the impending resolution of this goal as highly salient. Helming and colleauges (2014) offer a related explanation, drawing upon the well-established finding that children are highly motivated to engage in spontaneous helping behavior (Warneken & Tomasello, 2007, 2009). They suggest that children in the change-of-location FBT would treat the fact that Sally has an unfulfilled goal as highly salient, and would be very motivated to help her fulfill that goal. Consequently, children infer that the experimenter must be indirectly asking them to help Sally; they thus interpret "Where will Sally look for her marble?" as "Where *should* Sally look for her marble? Let's *help* her find it!" They respond by giving the most helpful answer possible – namely, by indicating the actual location of the ball. In other words, children fail to judge that the doxastic interpretation of the experimenter's question is correct because they judge it far more likely that the experimenter is concerned with (what they see as) the most salient feature of the

situation: that Sally needs help. Thus, when children hear the experimenter ask, "Where will Sally look for her marble?" according to the pragmatic development account, they think, *the experimenter wants me to help Sally find her marble.*

This explanation also enables us to make sense of the fact that children tend to do slightly better on FBT tasks in which the change of location is the result of a deliberate deception (Chandler et al., 1989; Hala, Chandler, & Fritz, 1991; Wellman et al., 2001). In such a context the child may find it less likely that the experimenter (who is also the deceiver) would be inviting the child to undo his deception, which in turn raises the probability that the experimenter is doing something other than inviting the child to help. It also helps to explain why younger children succeed on false-belief tasks that use helping as a dependent measure: when children's inclination to help is exploited by the false-belief task design, and does not interfere with it, children's early false-belief competence is on full display (D. Buttelmann et al., 2009, 2014; Southgate, Chevallier, & Csibra, 2010).

If children's failures on the FBT are due in part to the fact that they do not see doxastic facts as conversationally relevant, this would suggest that manipulations that raise the salience of these facts should improve performance. A number of findings in the literature support this prediction. As we saw, the Lewis et al. (2012) and Hansen (2010) studies showed that the presence of contrastive information can heighten the salience of belief facts and thus trigger a doxastic interpretation. Asking a child where an agent will look *first* also leads to improved performance (Siegal & Beattie, 1991; Surian & Leslie, 1999); perhaps this added specificity simply restricts the range of plausible interpretations, forcing the child to consider the doxastic one more

carefully. Asking a child to play out a character's actions using a toy rather than directly querying them about the character's actions (i.e. the "Duplo Task") also helps, perhaps because the play-acting activity naturally leads the child to treat the character's beliefs as contextually relevant (Rubio-Fernández & Geurts, 2013); unlike traditional FBTs, the Duplo Task presents children with a situation in which showing their understanding of beliefs actually makes sense from their point of view. All of these manipulations, whether they involve the verb 'think' or not, seem to change the features of the situation in a way that makes children regard doxastic facts as more salient, enabling them to demonstrate their knowledge of the character's beliefs.

4.2. Social experience and the FBT

Here, the importance of social experience for understanding the relevance of belief facts becomes clear: children who have had more opportunities to observe and participate in conversations about beliefs seem to be better attuned to the conversational relevance of psychological facts. They may, for instance, gradually encounter more situations in which non-doxastic interpretations of speech acts fail to explain speakers' behavior, forcing them to entertain alternative, doxastic interpretations. In this manner, children may come to learn that the concept of *belief* that they deploy to interpret the behavior of others is also regularly implicated (either explicitly or implicitly) in everyday speech, especially in contexts involving diverse beliefs (Lewis et al. 2012) and false beliefs (Papafragou et al., 2007). This newly acquired knowledge prompts children to adjust their prior expectations about the potential relevance of belief-facts when drawing inferences about speaker meaning. They are then better able to disambiguate indirect and attributive uses of 'think,' and,

most importantly for our current discussion, accurately interpret experimenter queries in the FBT.

This experience could be achieved via exposure to maternal "mind-minded" conversation (Ruffman et al., 2002), interactions with older siblings (Perner et al., 1994; Ruffman et al., 1998), or various forms of explicit training (Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003). Notably, the absence of these experiences would lead to corresponding delays on FBTs. Late-signing deaf children, for instance, are not exposed to belief discourse until primary school, and consequently they show delays in explicit false belief performance (Wellman et al., 2011); yet, when they are exposed to theory of mind-based training interventions, they rapidly improve (Wellman & Peterson, 2013). The first cohort of Nicaraguan signers did not even possess mental state vocabulary when Pyers and Senghas (2009) first tested their explicit false belief competence, which they systematically failed. Several years later, after being exposed to the mental state vocabulary of the second cohort, their performance markedly improved. According to the pragmatic account, what developed in the interim was not a new set of concepts; rather, it was their sensitivity to the contextual factors that rendered beliefs conversationally salient. For the late-signing deaf-children, their general deficit in linguistic experience meant that they lacked crucial experience with belief discourse; Wellman and Peterson's intervention succeeded in compensating for this deficit. For the first-cohort Nicaraguan signers, the language itself was impoverished with respect to mental state terms, which resulted in impoverished experience with belief discourse.

These findings, which resist explanation under accounts that appeal solely to the executive demands of the FBT to explain systematic failures, are convincingly explained under the pragmatic development account. But more importantly, they point to the specific importance of experience with mental state discourse in improving children's performance on the FBT. Such experiences provide a developmental scaffold for the ability to understand when psychological facts are conversationally relevant.

At this point, one might object that while the pragmatic development account is well-equipped to explain experience-dependent individual differences in FBT-performance, it may seem less obvious that it can explain why most children suddenly pass this task around 4-and-a-half years of age. This sharp developmental shift was convincingly explained by the processing-load account, as it seemed indicative of a biologically-based, maturational change to a child's executive abilities. But now that the processing-load account has been shown to be inadequate (for the reasons given in Section 2), it is not clear that the space left in its wake can be filled by appealing solely to a child's experiences with belief discourse.

The strength of this objection depends upon the claim that there is a sharp shift in performance on the FBT around 4.5 years of age. But as we've already seen, the precise timing of this shift is in fact quite variable: children with more experience with belief discourse pass the FBT slightly earlier, while children with less experience pass it slightly later; children with significantly less experience (e.g. late-signing deaf children and first-cohort speakers of NSL) pass it significantly later. When we include populations from a broader range of cultures, the age of success on

the FBT varies still further: for instance, children from Hong Kong do not pass the task until they are 6 (Wellman et al., 2011), and Samoan children do not pass the FBT until around 8 years of age (Mayer & Trauble, 2012). Moreover, even in culturally homogeneous, Western populations, the shift in children's performance is not particularly abrupt: at 30 months of age, the rate of success on the task is 20%; by 40 months, it rises to 50%; by 56 months, it rises to 74.6% (Wellman et al., 2001). There is, in other words, a gradual, linear change in the rate of FBT success, the timing of which varies substantially in different populations. This broad pattern is consistent with the hypothesis that FBT performance is related to gradually increasing exposure to belief-discourse, which itself may vary on individual and cultural bases.[9] Thus, the pragmatic development account is fully able to explain why children pass the FBT when they do, whether or not this happens to occur at 4.5 years of age.

4.3. Desire discourse

The pragmatic development account also helps us understand another major developmental finding in the theory of mind literature, namely that children consistently succeed on verbal tasks that implicate the concept of desire well before those that involve false beliefs (Hadwin & Perner, 1991; Rakoczy et al., 2007; Wellman & Woolley, 1990). Explaining these findings has proven challenging for nativists, who hold that basic conceptual understanding of both belief and desire

[9] While some authors have appealed to deep cultural differences between Western and Eastern societies (e.g. individualist versus collectivist value systems) in order to explain cultural variation on theory of mind tasks, (Liu, Wellman, Tardif, & Sabbagh, 2008; Shahaeian et al., 2011), others have proposed that much of this variation may be due to factors that cross-cut such cultural divides: for instance, Hughes et al. (2014) suggest that the age at which children begin primary schooling, and the experiences that they gain in formal pedagogical contexts, may explain differences in FBT performance among various child populations, both in Western and Eastern countries (Hughes et al., 2014).

emerge in the first year of life. Leslie and colleagues (Leslie et al., 2004) have argued that desire-based tasks are less demanding on a child's executive resources than FBTs; however, Rakoczy et al. (2007) have shown that the gap between desire and false belief persists even when both types of task are matched for logical complexity.

An initial prediction of the pragmatic development account is that this phenomenon is likely to have its roots in children's conversational experiences, and indeed, there is reason to believe that this is the case. As I argued in section 3, both our explanations and descriptions of behavior tend to omit any overt reference to beliefs and refer only to desires, which leads references to desires to be roughly twice as frequent as talk of uses of 'think';[10] thus, children have a much greater input for desire discourse than for belief discourse (see also Smiley & Huttenlocher, 1989; Taumoepeau & Ruffman, 2006). Moreover the frequency of indirect uses of 'think' makes the input for belief discourse fairly noisy, whereas 'want' does not seem to pose the same kinds of pragmatic difficulties.[11] One would expect, then, that proficiency with desire-discourse would precede proficiency with belief-discourse, as the input for the former would be both greater and more easily interpretable than the input for the latter. Thus, according to the current account, children succeed on tasks involving desire before they succeed on tasks involving belief because desire discourse lacks the pragmatic difficulties posed by belief discourse.

---

[10] The asymmetric roles of beliefs and desires in folk psychological explanation is itself a fact in need of some explanation. Steglich-Petersen and Michael (2015) have recently argued that this is due to the fact that one may substitute one's own beliefs into most folk psychological explanations and still have them make sense, but that the same is not true of our desires; thus, we must make overt reference to desires in our folk psychological explanations because this is information that cannot be presupposed in a coherent explanation of behavior.

[11] 'Want' *can* be used imperatively (e.g. "Do you want to cut that out?" really means "Cut that out!") But given that desire discourse is also more frequent than belief discourse, genuine attributive uses of 'want' are likely to be common enough that it would not pose a comparable learning challenge.

5. A problem case: Call & Tomasello (1999)

One set of findings that seems to raise doubts about the pragmatic development account is reported by Call and Tomasello (1999), who developed an entirely non-verbal change-of-location FBT. In this task, two groups of children (with mean ages of 4 and 5 years) were asked to sit in front of a large rectangular barrier. Behind the barrier were two identical boxes. One experimenter, the Hider, would place a sticker inside a box while it was behind the barrier, where the child could not see it. Seated behind the Hider was the Communicator, whose job was to point to the box containing the sticker after the barrier was raised. The child was told that her job was to point to the box containing the sticker, which she would then be able to keep.

On the crucial false-belief trials, the Communicator would watch the Hider place the sticker in one of the boxes, and then briefly left the room. During the Communicator's absence, the Hider would switch the locations of the two boxes. The Communicator would then return and point to the location in which the sticker was originally hidden prior to the switch. In order to pass this task, children had to recognize that the Communicator had a false belief about the location of the sticker, and then ignore the Communicator's pointing gesture. Children's performance on this task was then compared to their performance on a standard FBT.

Call and Tomasello found that children's performance on this non-verbal task was highly similar to their performance on verbal versions of the FBT, with performance on both types of task improving with age. In other words, the younger children tended to fail both tasks, while the older children tended to pass them. These results are problematic for the pragmatic development account for two reasons. First,

the task is non-verbal, and made use of simple pointing gestures, which children understand well by this age (Behne, Carpenter, & Tomasello, 2005; Liszkowski, Carpenter, & Tomasello, 2007). This seems to eliminate the possibility that children were responding to an indirect speech act. Second, children of both age groups clearly seemed to understand that their goal in this task was to collect the stickers for themselves. This makes it unlikely that children in the crucial false-belief trial thought that they were being invited to help the ignorant communicator. Thus, both the pragmatic development explanation of children's misunderstanding in the standard FBT, and the explanation for their systematic errors in that task are off the table.

However, there is reason to believe that the correlation between children's performance on this non-verbal task and the standard FBT is illusory. While Call and Tomasello's design did reduce the pragmatic demands of the FBT, it also increased its inferential complexity and executive demands. In an ordinary change-of-location FBT, children must simply recall where the agent thinks the object is in order to give a correct response. In this task, in contrast, the child must 1) track the visible displacement of the sticker in the original hiding phase; 2) track the invisible displacement of the sticker during the switching of the boxes, 3) remember which parts of the task the Communicator was present for, and 4) ignore the Communicator's advice when she returned on the basis what the she remembers of 1)-3). Already, this exceeds the processing demands of an ordinary FBT.

In addition to this added executive burden, the child must be capable of reasoning with non-specific, quantified belief-attributions. Because the child does not know

where the sticker is, she must not represent, "The Communicator knows the sticker is in Box A," but rather, "The Communicator knows which box the sticker is in." Then, when she sees the boxes switched, she must infer, "Whichever location the Communicator thought the sticker was in, it is now in the other one." Then, when the Communicator points to one box, the child must select the other. This kind of reasoning is vastly more complex than what is required in the FBT. Thus, it appears that the above task design has simply traded pragmatic challenges for greater inferential complexity executive demands.

Call and Tomasello do attempt to head off this line of criticism by administering a series of control tasks, testing children's ability to track the visible and invisible displacements of the sticker, as well as the child's ability to ignore the Communicator. The authors thus argue that the children were capable of surmounting all of the executive and inferential demands of the task. Crucially, however, the study only controlled for each of these factors *independently* of one another, whereas the real challenge of the task would have been coping with all of these demands *simultaneously*. The study thus failed to adequately control for the executive demands and inferential complexity of the task.

6. Predictions

One could empirically distinguish between the pragmatic development account and the standard nativist "processing load" account in several ways. The processing load account points to the information-processing demands of the FBT itself to explain younger children's failures, and predicts that if we reduce these demands, children's performance should improve. In the past, this approach has been successful, and

several authors have developed simplified, less executively demanding versions of the FBT that children are able to pass before their fourth birthdays (Rubio-Fernández & Geurts, 2013, 2015). The pragmatic development account, in contrast, points to the conversational context of the FBT as a crucial determiner of performance. It predicts that if children are better able to grasp the fact that in this context they are supposed to attend to belief facts, then their performance on a standard FBT should improve – even if *no change* is made to the task's immediate information-processing demands. Thus, where the processing load account would predict that contextual manipulations that leave the basic executive demands of the FBT unchanged should have no effect on performance, the pragmatic development account predicts that these manipulations should lead to improvements in performance.

This contextual manipulation could be achieved using a between-subjects design with children aged 3.5 to 4 years (i.e. on the cusp of passing the FBT). Both the experimental group and the control group would complete a standard, change-of-location FBT as the dependent measure. But prior to completing the task, the experimental group would engage in a pre-test familiarization activity that would serve to heighten the saliency of beliefs. One might achieve this effect by having children answer a number of questions that draw a contrast between knowledge and belief. One way to do this would be to base the activity on Wellman and Peterson's "thought bubble" intervention, originally used with deaf children (Wellman and Peterson, 2013). Unlike prior research that used multiple training sessions to improve children's performance on the FBT, the pragmatic account predicts that these manipulations should have an immediate effect.

Alternatively, one could use a contextual manipulation to *diminish* the saliency of alternative interpretations of the FBT query – for instance, by making the prospect of helping the agent seem undesirable (Helming et al., 2014). This could be achieved by borrowing from the design of Vaish, Carpenter, and Tomasello (2010). These authors found that three year-old children would selectively avoid helping an agent if they previously saw that agent intentionally harm someone else. If children's incorrect response in change-of-location FBTs is in fact a helping response, then seeing an agent commit an intentionally harmful act prior to completing the FBT should also diminish their inclination to give this response.

One could also use the same type of contextual manipulation to cause children who we would expect to pass the FBT (e.g. children aged 4.5 to 5) to fail the task. This could be achieved by making the needs of the agent especially salient and/or by emphasizing the prosocial nature of the agent, causing the motivation to help to override the still-fragile doxastic response. This manipulation could be implemented by showing children an agent with a false belief who has just been the victim of an unfair resource allocation. Since we know that children at this age are highly motivated to rectify such inequalities (Li, Spitzer, & Olson, 2014), we might expect them to use the context of the FBT as a chance to rectify the inequality, and thus give the helpful but incorrect response.

Crucially, all of these designs would include a standard FBT task, free of simplifying manipulations. However, one might also build on the last suggestion by taking a simplified FBT that younger children normally pass (e.g. Rubio-Fernandez and Geurts, 2013, 2015), and using a contextual manipulation to make it harder. The

pragmatic development account predicts that even if the processing demands of the task have been sufficiently reduced, a contextual manipulation emphasizing the helping motivation should cause these children to fail. Something like this might explain one of the results of Rubio-Fernandez and Geurts (2015). In their Experiment 2, they administered a simplified version of the FBT – the "Duplo task" – that three-year-olds have been known to pass with little difficulty, but with one modification: they made sure to *mention the desired object.*[12] This caused three-year-olds overwhelmingly to fail the task. The authors attribute children's failure in this manipulation to the fact that mentioning the object made it more salient to the child, disrupting the child's perspective-taking in the process. But the pragmatic development account offers a different interpretation: drawing the child's attention to the banana *qua* object of desire triggered their motivation to help, causing them to lead the character to the bananas' correct location instead of acting out the appropriate behavior.

An alternative to the contextual manipulation strategy would be to modify the immediate FBT context in a way that would increase the salience of belief facts. This kind of approach would probably make it more difficult to tease apart pragmatic factors from executive demands, since it could be claimed that any resulting improvement in performance might be due to a reduction in processing demands, rather than anything to do with the salience of belief facts. But this sort of obstacle

[12] In the Control Question condition, the experimenter first asks, "Where is the banana now?" and then proceed with the test question. In the Goal condition, the experimenter says, "Now Lola is very hungry and wants the bananas" (Rubio-Fernández & Geurts, 2015, p. 10). Of the two conditions, the Goal condition is most clearly explained by the helping interpretation; however, the Control Question condition is amenable to it as well. Merely mentioning the bananas may be enough to trigger the helping motivation.

could be overcome by following the example of Lewis et al. (2012). Recall that Lewis and colleagues were able to improve children's performance on a truth-value judgment task for belief reports by *adding* an additional seeker with contrasting beliefs (see Section 3.2). This manipulation would have also increased the processing demands of the task – reading two minds is harder than reading one – but this did not seem to hurt children's performance. One could implement a similar manipulation in a more traditional FBT design. In such a task, children would be presented with a hide-and-seek scenario; in the 1-seeker condition, children would see just one seeker with a false belief about the location of the hider; in the 2-seeker condition, children would see two seekers, one with a false belief about the hider's location, and the other with a true belief. In both conditions, the test question would be, "Where will [the seeker with the false belief] look for [the hider]?" The pragmatic development account would predict that having two seekers as opposed to one should improve performance, whereas the processing load account would predict that it would lead to a decrease in performance.

7. Conclusion

In this paper, I've illustrated how belief discourse poses substantial challenges for young children that have nothing to do with whether or not they possess the concept of belief. This highlights a new way of interpreting the relationship between children's early social experiences and their performance on FBTs: even children who are able to represent false beliefs must still learn from their social environment how and when belief facts are implicated in conversation before they are able to pass the FBT. Thus, if a child's social environment is enriched or impoverished with

respect to belief discourse, this will affect when she passes the FBT. The pragmatic development account thus provides the theory of mind nativist with a framework for accommodating a wide range of variation in FBT performance brought on by differences in individuals' social experiences, as well as set of empirical predictions for testing and extending that framework and enriching our understanding of theory of mind development.

Chapter 3: Spontaneous mindreading: A problem for the two-systems account[13]

1. Introduction

For decades, social cognition research has been dominated by the idea that we navigate the social world by attributing mental states to other individuals in order to predict and explain their behavior – the ability known as "theory of mind" or "mindreading" (Carruthers, 2013; Fodor, 1992; Goldman, 2006; Nichols & Stich, 2003). This approach to social cognition has been quite fruitful, and has yielded an immense body of empirical knowledge about the development of our social cognitive abilities and their neural underpinnings (Baillargeon et al., 2010; Saxe & Kanwisher, 2003; Wellman, 2014). But philosophers and psychologists are nevertheless divided over how great a role these folk-psychological concepts actually play in our everyday lives. While many continue to assume that the mindreading paradigm is basically sound, others have suggested that it is deeply flawed as an account of our ordinary socio-cognitive abilities, and must be radically re-thought.

One of the most compelling skeptical arguments about mindreading draws our attention to the unbounded scope of paradigmatic folk-psychological inferences (Bermudez, 2003; Morton, 1996; Zawidzki, 2013). This argument begins with the idea that belief-formation itself is a holistic, unbounded, "isotropic" process (cf. Fodor (1983)). Our actions can be informed by an indefinitely wide range of beliefs and desires. For instance, when I decide to take the metro rather than drive, I may do

[13] This chapter was originally published as Westra (2016b). It has been reprinted here with permission from Springer, copyright license # 4067640831758.

so because I believe that taking the metro is better for the environment, and I desire to make environmentally-friendly choices; it may also be because I think that parking is expensive, and I wish to save money; or I may believe that I am being followed, and I wish to lose my pursuers on the crowded metro platform. There are, in other words, indefinitely many different folk-psychological ways to rationalize a particular action. When interpreting other people's actions, the argument goes, we are faced with the daunting task of sifting through this immense space of possible mental causes, and abductively inferring which belief-desire set best explains the action in question. Such a process would no doubt be incredibly demanding and effortful, and would place heavy demands on executive systems like working memory – that is, if the task were not completely intractable. It thus seems highly unlikely that we engage in this kind of inference during our everyday social interactions, which occur at a very rapid pace.

Motivated by these and other skeptical concerns, a number of theorists have proposed alternatives to mindreading in order to explain our everyday social-cognitive abilities. Some have suggested that we gain knowledge of mental states via automatic, perception-like processes (Gallagher, 2008). Others have argued that we draw on folk-psychological narratives and social norms to predict behavior, rather than belief-desire inferences (Hutto, 2012). Still others have suggested that many of our socio-cognitive abilities may be subserved by a combination of low-level associations between perceptions of behavior and domain-general attentional processes (Heyes, 2014b). It has even been suggested that these abilities are parts of dynamical systems that emerge during social interactions with multiple agents, and

thus cannot be explained in individualistic terms and all (De Jaegher & Di Paolo, 2007).

It is worth emphasizing the broad-reaching, often radical consequences of these anti-mentalistic proposals. Mindreading is widely believed to be central to many uniquely human social practices: linguistic communication (Grice, 1991; Wilson & Sperber, 2012), moral judgment (Mikhail, 2007; Thomson, 1976; Young, Cushman, Hauser, & Saxe, 2007), joint action (Bratman, 1992; Tomasello, Carpenter, Call, Behne, & Moll, 2005) and establishing new social conventions (D. Lewis, 1969). Our grasp of the psychological underpinnings of these activities hinges on the view that mindreading is a cornerstone of social cognition. If we abandon the mindreading paradigm, then our theories about these important human activities must also be re-thought.

In the context of this dispute over the scope of theory of mind in our everyday social lives, the two-systems account of mindreading (Apperly and Butterfill 2009; Apperly 2011; Butterfill and Apperly 2013) seems to offer something of a middle ground: on the one hand, it adheres to the basic idea that some form of mindreading is pervasive in our everyday lives. But on the other hand, two-systems theorists agree with mindreading skeptics that the attribution of "full-blown" propositional attitudes such as beliefs is generally quite slow and effortful, places heavy demands on attention and working memory, and likely requires fairly advanced linguistic abilities. Because it is so demanding, proponents of the two-systems account agree that this form of reasoning is unlikely to contribute to many of the ordinary social practices that are often associated with it. When it does, it is likely scaffolded by some of the

very processes proposed by anti-mindreading theories, such as social norms and narratives (Apperly 2011).

And yet, two-systems theorists also maintain that we are nevertheless equipped with an innately-channeled, automatic mindreading system that is constantly active in the presence of other agents. However, the representational capacities of this system, according to the two-systems account, fall well short of the kind of unconstrained belief-desire reasoning typically associated with mindreading. Instead, this "implicit" mindreading system is said to employ a limited set of quasi-mentalistic, mainly extensional concepts and inference rules that allow us to roughly predict behavior. Because of its limited representational capacities, this system exhibits a number of signature limits that distinguish it from genuine, "explicit" mindreading. Specifically, the implicit mindreading system is said to be insensitive to the fact that agents represent the world under a particular *mode of presentation*. For example, this system could never predict that Lois Lane would be surprised to see Clark Kent fly, even if she knew that Superman can fly, and that Superman is Clark Kent, because the implicit system would be insensitive to the fact that a single individual can be represented in a number of different ways.

Thus, according to this view, humans possess two "systems" for mindreading: the early-developing, automatic "implicit" system, and the later-developing, slow and effortful "explicit" system. Initially, human infants start out with just the implicit mindreading system, and as a result, their mindreading abilities are subject to its signature limits. As they get older, acquire language, and gain social experience, children develop explicit mindreading abilities, which start to emerge after their

fourth birthday. Ultimately, these two systems exist side by side in adulthood, producing distinct types of mental state judgments in parallel to one another, creating a dissociation between implicit and explicit forms of mindreading (Low & Perner, 2012).

The main goal of this paper is to offer a critique of the two-systems account of mindreading. Specifically, I will be arguing that the two-systems account is unable to accommodate the extant empirical evidence on one, very central type of mental-state attribution: the attribution of perceptual states or "perspective-taking." I'll further argue that these problems generalize to other aspects of theory of mind, and thus seriously undermine the two-systems account. What emerges in its place is a picture of mental-state attribution that lies somewhere in between the automatic, rigid information processing of the implicit mindreading system, and the slow, effortful reasoning of the explicit system. The evidence that I will discuss suggests that even "full blown" forms of mental-state attribution can be both fast and flexible, and that "implicit" forms of mindreading can be highly flexible and context-sensitive. This combination of speed and flexibility is achieved via the coordinated integration of domain-specific mindreading strategies with goals, attention and knowledge stored in long-term memory.

But while the main target of this paper is the two-systems account, this critique has broader implications for mindreading skeptics as well. This is because a key flaw in the two-systems account is that it accepts the skeptic's claim that genuine mindreading must be slow and cognitively effortful. A proper understanding of the underlying processes that enable mindreading shows that this is a mistake. Thus, the

picture of mindreading that emerges from my critique of the two-systems account can also serve as a reply to the skeptic: we should not abandon the mindreading paradigm so quickly.

In the second section of this paper, I will discuss the general theoretical motivations for the two-systems account. Then, in section 3, I will introduce the notion of perspective-taking as it occurs in the social cognition literature, and explain how the two-systems account purports to explain perspective-taking phenomena. In sections 4 and 5, I will argue that the evidence from perspective-taking undermines key claims about the implicit and explicit mindreading systems, respectively. In section 6, I will show how the problems from perspective-taking generalize, and ultimately undermine the two-systems account as a whole. In section 7, I'll discuss the fast, flexible conception of mindreading that emerges from my critique, and how it can serve as a bulwark against theory-of-mind skepticism.

2. Why two systems?

The motivation for proposing two systems for mindreading, its proponents argue, becomes especially clear when we consider the kinds of properties that human mindreading must possess in order to successfully navigate ordinary social interactions. First, our mindreading abilities must be very fast and efficient, in order to keep up with the pace of ordinary behavior. Second, they must also be representationally flexible, since we need to be able to attribute an indefinite range of attitude contents to others in order to make sense of the complexity of human behavior. The problem, according to two-systems theorists, is that,

> [T]here is a tension between the requirement that mindreading be extremely flexible on the one hand, and fast and highly efficient on the other. Such characteristics tend not to co-occur in cognitive systems, because the very characteristics that make a cognitive process flexible – such as unrestricted access to the knowledge of the system – are the same characteristics that make cognitive processes slow and effortful. Instead, flexibility and efficiency tend to be traded against one another. This trade-off is reflected in Fodor's distinction between "modular" versus "central" cognitive processes. (Apperly, 2013, pp. 73–74)

Thus, human beings need at least two mindreading systems because no single system could be *both* efficient and representationally flexible. According to this view, we rely on the fast and inflexible system when we need to rapidly anticipate what others will do, while we turn to the slow, flexible system when we need to carefully reflect on their specific beliefs. Thus, the reason the implicit system is unable to represent "full-blown" propositional attitudes is because these are thought to place heavy demands on working memory, which is slow but representationally flexible (Butterfill & Apperly, 2013).[14] The implicit system gains its speed and efficiency from the fact that it can circumvent these forms of reasoning, and rely instead on a strictly limited set of quasi-psychological concepts and inference rules to automatically generate

[14] Why is representing "full-blown" propositional attitudes so demanding? Butterfill and Apperly write:

> On any standard view, propositional attitudes form complex causal structures, have arbitrarily nestable contents, interact with each other in uncodifiably complex ways and are individuated by their causal and normative roles in explaining thoughts and actions…. If anything should consume working memory and other scarce cognitive resources, it is surely representing states with this combination of properties. (Butterfill & Apperly, 2013, pp. 609–610)

See Carruthers (2015c) for a critique of this argument.

rough-and-ready predictions about behavior. But when accurate behavioral prediction means factoring in the *way* that an agent represents a particular state of the world, this system ought to make systematic errors. The explicit system, meanwhile, should be able to accommodate these cases; but this processing will inevitably be slow and effortful, and always goal-dependent.

The purported properties of the implicit mindreading system appear to derive from its modularity. In particular, Apperly (2010) emphasizes the essential role that *informational encapsulation* would play in the two-systems architecture. An informationally encapsulated, modular system could permit us to circumvent the need for effortful uses of working memory in many social interactions, thus rendering mental-state attribution fast and efficient, and even possible for young infants and non-human animals. However, such a system would also be representationally limited, due to its lack of access to working memory and stored knowledge. In other words, the tension between the need for flexible and efficient mindreading that motivates the two-systems proposal is explained by the trade-offs inherent in a modular, informationally encapsulated architecture.[15]

Moreover, Apperly (2010) suggests that informational encapsulation may provide part of the solution to the challenge raised by mindreading skeptics mentioned in the introduction. He argues that a modular, informationally encapsulated system could impose "hard constraints" on the scope of our folk-

[15] There are a number of other well-known modularist approaches to theory of mind (Fodor, 1992; Leslie et al., 2004; Scholl & Leslie, 1999); however, these accounts tend not to sharply distinguish between implicit and explicit mindreading systems, as the two-systems theorists do. Although a discussion of these views is beyond the scope of this paper, it is likely that many of the arguments to come that are directed at the two-systems account will also pose challenges for them as well.

psychological inferences, thus limiting the need for complex, abductive reasoning. By restricting the range of possible inputs that it could process, and by sharply delimiting the kinds of inferences that could be made on the basis of those inputs, an encapsulated, implicit mindreading system offers us a way to render mental-state attribution computationally tractable. Thus, the notion of informational encapsulation seems to provide the two-systems account with both a basic architectural framework and a potent theoretical justification.

Problematically for the two-systems view, there is growing consensus among cognitive scientists that perceptual systems – the paradigms of modularity (Fodor, 1983; Pylyshyn, 1999) – are not, in fact, informationally encapsulated.[16] Instead, we find that abstract, conceptual knowledge "penetrates" even the earliest, most rapid stages of visual processing. For instance, there is evidence that feedback signals from inferotemporal conceptual areas impact processing in the visual cortex just 100ms following stimulus onset, well before the onset of endogenous attention (Wyatte, Jilk, & O'Reilly, 2014). Similarly, Moshe Bar and colleagues have shown that conceptual information in the orbitofrontal cortex gets applied to rapidly transmitted, low spatial-frequency visual information, which is then projected back to mid-level and high-level visual processing areas 50ms before object-recognition takes place (M. Bar et al., 2006; Chaumon, Kveraga, Barrett, & Bar, 2014). There is also EEG evidence that linguistically encoded categorical distinctions (e.g. the lexical distinction between light and dark blue in modern Greek) can penetrate pre-attentional, pre-conscious processing in the visual cortex as early as 200ms after stimulus onset (Thierry,

[16] For a recent review of this topic, see Ogilvie and Carruthers (2016).

Athanasopoulos, Wiggett, Dering, & Kuipers, 2009). In short, there appear to be a number of pathways by which conceptual information stored in long-term memory can penetrate even paradigmatically modular systems, such as early visual processing.

The fact that vision is unencapsulated tells us something important: the trade-off between speed and representational flexibility is not mandated by our cognitive architecture.[17] But just because certain aspects of perceptual processing may be unencapsulated, it does not follow that there are no genuinely encapsulated systems. For instance, the analogue magnitude system, which served as a model for the implicit mindreading system (Apperly & Butterfill, 2009), may well be impenetrable to goals and information stored in long-term memory (Feigenson, Dehaene, & Spelke, 2004). However, the question still arises: is fast, efficient mindreading truly informationally encapsulated, like analogue magnitude reasoning? Or is it more like early vision, and capable of using both top-down and bottom-up information to rapidly and flexibly interpret the social environment?

3. The case of perspective-taking

A key test-case for the claim that the implicit mindreading system is truly encapsulated is the component of theory of mind known as "perspective-taking," which consists in the ability to represent what other agents see. In the empirical literature on the subject, there is a well-established distinction between two different "levels" of perspective-taking, which captures two different ways in which an

[17] In their own critique of the two-systems view, Christensen and Michael give a number of examples of well-studied cognitive systems that also succeed in achieving both flexibility and efficiency without the need for strong encapsulation, including the orbitofrontal cortex, the mid-level visual system, and language comprehension (Christensen & Michael, 2015).

organism might represent the visual perspective relation. "Level-1" perspective-taking consists in the ability to represent *what* another agent can see. Level-1 perspectives are construed as external, spatial relations that hold between agents and objects in their environments. This kind of relation depends primarily on environmental factors, such as an unobstructed line-of-sight, lighting, and distance. To be a Level-1 perspective-taker thus consists in representing whether this external relation is present or absent, and forming appropriate expectations about behavior on this basis. For instance, a Level-1 perspective-taker would not expect an agent wearing a blindfold to reach towards a goal object in front of her, because the blindfold interrupts her line-of-sight.

"Level-2" perspective-taking, in contrast, appears to be uniquely human. It involves representing the *way* that other agents see the world. Rather than a direct relation between agents and their environments, the Level-2 perspective relation holds between agents and representational contents; however, it depends upon some of the same environmental factors as Level-1 perspective-taking, such as line-of-sight. The key difference between Level-1 and Level-2 perspective-taking is that only the latter is sensitive to the representational, aspectual nature of vision.

To illustrate, imagine that you and a partner are seated opposite one another at a table, and lying flat upon the table is a screen with the numeral "9" on it. In the purely extensional, Level-1 sense, you would both see the same thing: "9". But in the intensional, Level-2 sense, you would each see something different: while you would see the numeral *as* the number nine, your partner would see it *as* the number six. In other words, the more complex Level-2 relation permits us to track differences in

mode of presentation.

In humans, Level-1 perspective-taking abilities emerge fairly early in development, and are even present in infancy (Luo & Johnson, 2009; Masangkay et al., 1974; Moll & Tomasello, 2006). A number of non-human animal species are also capable of Level-1 perspective-taking, including corvids, canines, and great apes (Bräuer, Call, & Tomasello, 2004; Bugnyar, Reber, & Buckner, 2016; Josep Call & Tomasello, 2008). The ability to represent Level-2 perspectives seems to emerge somewhat later in childhood, after the fourth year of life – the same age when children pass the standard false belief task. (Flavell et al., 1981; Low et al., 2014; Surtees et al., 2012). For this reason, Level-2 perspective-taking is said to signal children's acquisition of a representational theory of mind (Rakoczy, 2015).

Beyond its comparative and developmental applications, the Level-1/Level-2 distinction has also been invoked to describe adults' perspective-taking abilities. Specifically, it has been argued that representing Level-1 and Level-2 perspectives involve distinct cognitive processes (Michelon & Zacks, 2006; Qureshi et al., 2010; Samson et al., 2010; Surtees et al., 2012; Surtees, Samson, et al., 2016). Level-1 perspective-taking appears to be very rapid, places relatively few demands on executive resources, and seems to employ a simple line-of-sight heuristic. Level-2 perspective-taking, in contrast, appears to be slow, places heavy demands on working memory, and employs a kind of embodied mental rotation procedure (Surtees, Apperly, & Samson, 2013a).

The distinction between Level-1 and Level-2 perspective-taking thus seems to offer a clear-cut case of the dissociation between the implicit and explicit

mindreading: Level-1 and Level-2 perspective-taking each possess developmental and cognitive profiles that map fairly neatly onto the two mindreading systems. Accordingly, the two-systems account makes a number of specific predictions about perspective-taking that bear directly upon the issue of informational encapsulation. First, if the implicit system is truly informationally encapsulated, then Level-1 perspective-taking should be insensitive to the background knowledge of the perspective-taker. Second, if Level-2 perspective-taking truly places heavy demands on working memory, then we should expect it to operate in a goal-dependent fashion, and to be relatively slow and effortful.

To test the first prediction, Samson et al. (2010) created the "dot-perspective task." In this task, adult participants had to rapidly judge what either they or an avatar could see. Subjects were presented with a scene in which an avatar stood alone in a room facing a wall. In Consistent Perspective trials, black dots appeared on the wall that the avatar could see. In Inconsistent Perspective trials, some of the dots appeared on the wall that the participant could see, but the avatar could not. In the Self-task, participants had to judge how many dots they themselves could see; in the Other-task, they had to judge how many dots the avatar could see. Samson and colleagues found that people were much slower to respond and made more errors in the Self-task for Inconsistent perspective trials. Participants seemed to represent the avatar's Level-1 perspective even when it was irrelevant to their current goal, to the point that it interfered with their performance – exactly as the two-systems account predicted it would.

To test the prediction that the implicit system cannot represent Level-2 perspectives, Surtees et al. (2012) presented participants with another scene containing an avatar; but this time, instead of dots, the experiment used numerals displayed on a table in front of the avatar opposite the participant – the "number-perspective task." On Consistent Perspective trials, a numeral like '8' was displayed, which both the avatar and the participant saw the same way. In Inconsistent Perspective trials, a '6' or a '9' was presented on the table, which the participant and avatar would perceive differently; thus, this task required Level-2 perspective-taking abilities. As in Samson et al., participants completed both Self and Other tasks. Unlike in the Samson et al. experiments, the Inconsistent perspective of the avatar did not interfere with their response times on the Self-task. Participants appeared to only compute the other individual's perspective on the Other-task, when it was goal-relevant – once again, just as the two-systems account predicted.

While these results do seem to bear out the above predictions, a number of other findings in the perspective-taking literature are not so easily accommodated by the two-systems framework. In the next section, I will argue that we have good evidence that Level-1 perspective-taking is neither fully encapsulated nor truly automatic. In section 5, I will argue that Level-2 perspective-taking need not be slow and cognitively effortful.

4. Level-1 perspective-taking is unencapsulated: The argument from gaze-cueing

To see why the Level-1 perspective-taking evidence does not fully support the two-systems account, we need to consider another experimental paradigm that also aims to study implicit perspective-taking: gaze-cueing. Gaze-cueing tasks measure the

effects of shifts in the direction of a target's eye gaze or head on covert spatial attention – that is, changes in attention that happen *prior* to any overt forms of attention shifting, such as movements of the eyes or head (Posner, 1980). In gaze-cueing studies, subjects are presented with a task-irrelevant face in the center of a screen, with eyes that move either in one direction or another (Friesen & Kingstone, 1998; Hood, Willen, & Driver, 1998). Subjects then witness an object suddenly appear either on the same side as the direction that the face's eyes have "looked" (a congruent trial) or on the opposite side (an incongruent trial). The gaze-cueing effect occurs when subjects are faster to detect the object on the congruent side than the incongruent one. These effects are extremely rapid – on the order of 10-15ms - and are also specific to social stimuli (Kingstone, Tipper, Ristic, & Ngan, 2004; Ristic & Kingstone, 2005).[18] Thus, gaze-cueing seems like exactly the kind of effect that one might expect from the implicit mindreading system: it is extremely fast, unconscious, and tracks Level-1 perspectives.

If the implicit system were truly encapsulated, knowledge stored in long-term memory would not affect it. However, we know from a wide range of studies that gaze-cueing is in fact sensitive to background knowledge. For instance, Eva Wiese

[18] Since cueing effects can also be triggered by other kinds of directional stimuli, such as arrows (Ristic, Friesen, & Kingstone, 2002), some have suggested that this process might be the product of a domain-general covert orienting mechanism (Santiesteban, Catmur, Hopkins, Bird, & Heyes, 2014). However, these two types of cueing effects appear to have distinct cognitive, developmental, and neural bases. Specifically, gaze shifts appear to issue in a distinctly spatial cueing effect for the specific location where the eyes look, whereas arrows produce object-based cueing effects for any items that appear on the congruent side, regardless of their specific location (Marotta, Lupiáñez, Martella, & Casagrande, 2012). Further, while gaze-cueing effects appear even in extremely young infants (Farroni, Massaccesi, Pividori, & Johnson, 2009; Hood et al., 1998), cueing effects from other kinds of stimuli do not emerge until much later in development (Jakobsen, Frick, & Simpson, 2013). Finally, gaze-cueing, but not other kinds of cueing, produces activity in the superior temporal sulcus (STS), a neural region associated with social cognition (Ristic & Kingstone, 2005) (see also Michael and D'Ausilio (2015)).

and colleagues showed participants a robot-face cueing stimulus (Wiese, Wykowska, Zwickel, & Müller, 2012). In one experiment, they found that participants were much less likely to be cued by the gaze-shifts of the robot than those of a human face. However, in another experiment, participants were explicitly told that an experimenter was intentionally controlling the robot's gaze-shifts. In this condition, participants were just as likely to be cued by the robot-face as the human face. Thus, the presence of explicit, folk-psychological background knowledge about the stimulus affected whether or not partners were cued by an otherwise non-agentive stimulus.

Similarly, when a cueing stimulus is ambiguous, background knowledge about whether or not it is an intentional agent can modulate whether it produces a cueing effect. Ristic & Kingstone (2005) showed subjects an ambiguous stimulus, and told them that two eye-like shapes were either eyes or wheels on a car; they found cueing effects for the eyes condition, but not for the car condition. Even more strikingly, Terrizzi and Beier recently showed participants an unfamiliar entity and modulated whether or not, prior to the cueing trials, subjects saw another agent appear to interact with it in a contingent, seemingly social manner. They observed "gaze" cueing effects for the unfamiliar entity (even though it did not, in fact, possess eyes, but merely a presumed front and back) in the social interaction condition, but not in the non-social condition (Terrizzi & Beier, 2016).

Background knowledge about whether or not a human face can see also modulates the cueing effect. Teufel and colleagues showed participants images of a face wearing goggles; beforehand, subjects had the opportunity to handle a seemingly identical pair of goggles (Teufel, Alexis, Clayton, & Davis, 2010). However, one

group handled goggles with opaque lenses (such that the wearer would not be able to see through them), while another group handled goggles with transparent lenses. They found that only participants who handled the transparent goggles were cued by the head-movements of the stimulus. Thus, if participants knew that the face could not see, the cueing effect was attenuated.

Importantly, these studies always showed subjects in both experimental and control conditions *perceptually identical stimuli*; all they varied was the background knowledge that subjects had about what they were looking at. In other words, these studies provided a perfect test for informational encapsulation, and showed that gaze-cueing is not encapsulated after all. Thus, contrary to the two-systems account, background knowledge affects Level-1 perspective-taking.

One study from the two-systems group offers a potential avenue for them to respond to this point. Using the same stimuli as in the Samson et al. study described above, Qureshi et al. (2010) tested whether or not Level-1 perspective-taking would be affected by concurrent executive demands; according to the two-systems account, it should not. To do this, they used a dual-task interference design in which subjects had to complete the dot-perspective task while simultaneously tapping along with a recorded beat. They found that the cognitive load task did interfere with task performance, but this interference was similar for both the Self- and Other-tasks. While this finding might initially be interpreted as undermining the claim that Level-1 perspective-taking is truly an efficient process, the authors argued that the similar interference effects for both Self- and Other-trials showed that the tapping task did not interfere with the *calculation* of Level-1 perspectives as such, but rather with the

attentional *selection* of perspectives in general. According to this picture, the Level-1 perspective-taking process would involve two components: a perspective-selection process that places demands on domain-general attention, and a domain-specific, encapsulated mechanism for perspective-calculation.

Accordingly, proponents of the two-systems account could argue that all the gaze-cueing studies show is that the Level-1 perspective-*selection* process is unencapsulated from background knowledge, but that the perspective-calculation process is not. Thus, in cases when the gaze of a target face is known not to be indicative of genuine seeing, that perspective might not be selected by attention, and thus no perspective-calculation would occur. But it could still be maintained that Level-1 perspective-*calculation* is encapsulated, once a given perspective has been selected.

This distinction between selection and calculation enables the two-systems theorist to maintain that that there could be an encapsulated mechanism for Level-1 perspective-calculation. But at best, such a mechanism could only be one component of the system that performs the function of Level-1 perspective-taking. This is because perspective-selection also seems to be a necessary part of the perspective-taking process: in the absence of perspective-selection, no perspective-taking could take place. Thus, we could not ascribe the function of Level-1 perspective-taking solely to the perspective-calculation mechanism. If there is a "system" that is responsible for Level-1 perspective-taking, then it must also include whatever mechanism or mechanisms that accomplish perspective-selection – and these, it appears, are unencapsulated. Thus, while the "system" responsible for Level-1

perspective-taking might involve component parts that are informationally encapsulated, this does not change the fact that Level-1 perspective-taking as such *is* sensitive to background knowledge.

Moreover, acknowledging a role for domain-general attention in the perspective-taking process also undermines the claim that Level-1 perspective-taking is truly *automatic* – that is, if by "automatic" we mean a process that is mandatory, stimulus-driven, and goal-independent (Moors & De Houwer, 2006). This is because, more often than not, domain-general attention is goal-directed (Carruthers, 2015b). In paradigmatic instances of "top-down" attentional orienting driven by the dorsal orienting network, these goals are conscious. But attention can also be controlled by the ventral orienting network, which is sensitive to unconscious goals and motivations (Corbetta, Patel, & Shulman, 2008).[19] Thus, by acknowledging a role for attention in perspective-selection, two-systems theorists are opening up a space where goals might play a significant role in the Level-1 perspective-taking system.

Consistent with this possibility, other studies have shown that knowledge of the social group memberships of a target face, including its age, race, social status, and perceived threat all affect gaze-cueing (Chen & Zhao, 2015; Dalmaso, Pavan, Castelli, & Galfano, 2012; Pavan, Dalmaso, Galfano, & Castelli, 2011; Slessor, Laird, Phillips, Bull, & Filippou, 2010). In addition to interactions between the gaze-cueing mechanisms and long-term memory, these findings show that gaze-cueing is also

[19] Granted, attention can sometimes be "captured" in an automatic, goal-independent manner by environmental stimuli (E. I. Knudsen, 2011), and it's conceivable that Level-1 perspective-taking could likewise be the product of purely bottom-up processing. However, many of the gaze-cueing experiments cited above were able to perfectly control for such low-level effects by using perceptually identical stimuli in both experimental and control conditions. The factors that modulated Level-1 perspective taking in these experiments could not have been purely stimulus-driven.

sensitive to motivational factors: when a face is motivationally salient – for instance, because it belongs to a threatening out-group member – we preferentially allocate attentional resources in order to follow its gaze. However, when a face is not motivationally salient – say, because it belongs to a low-status in-group member – we do not preferentially attend to its gaze direction. In other words, Level-1 perspective-taking appears to be highly sensitive to our social goals.

Thus, the evidence from gaze-cueing seems to shows that Level-1 perspective-taking is neither wholly encapsulated, nor truly automatic. Of course, Level-1 perspective-taking is also not under explicit, top-down, conscious control. Rather, its information-processing profile seems to belong somewhere in between these two. It is better described as a "spontaneous" process: it is fast, efficient, and unconscious, but also sensitive to background knowledge and goals (Carruthers, 2015a). Notably, this kind of process does not quite fit with the descriptions of either the implicit or explicit systems. Instead, it seems to share attributes of both.

If this picture is right, and Level-1 perspective-taking is really spontaneous, rather than automatic, then why do subjects in the dot-perspective task represent the avatar's Level-1 perspective? This did, after all, conflict with their overt goal, and it is not obvious what else might have motivated participants to attend to its perspective. One possibility is that even though Level-1 perspective-taking is not genuinely automatic, we may possess a standing disposition to represent other agents' perspectives when doing so is cognitively efficient.[20] Given that what other agents

[20] Along similar lines, Fiebich & Coltheart (2015) suggest that which socio-cognitive procedure we use is determined by whether or not it will be cognitively effortful in a given context (Fiebich & Coltheart, 2015). (Thanks to an anonymous reviewer for bringing this reference to my attention).

can see tends to be behaviorally relevant, and that calculating Level-1 perspectives is not particularly demanding, such a disposition would be fairly adaptive in most situations. In practice, this might make Level-1 perspective-taking seem automatic in most situations, when in fact it is really motivation-dependent.

5. Level-2 perspective-taking can be fast and efficient

The argument from gaze-cueing suggests that Level-1 perspective-taking does not quite fit with the description of the implicit mindreading system as automatic and encapsulated. However, it leaves untouched the basic claim that Level-2 perspective-taking should be a slow, effortful, working-memory-based process. Thus, two-systems theorists may be willing to concede that Level-1 perspective-taking is more flexible than they initially supposed, but still argue that Level-2 perspective-taking, which involves "full-blown" propositional attitude attribution, must possess something like the cognitive profile of the explicit mindreading system.

Notably, Level-2 perspective-taking tasks almost always involve some kind of mental rotation (Flavell et al., 1981; Low et al., 2014; Surtees, Apperly, & Samson, 2013b), as this seems to be one of the most straightforward empirical methods for creating a dissociation between Level-1 and Level-2 perspectives. Problematically, mental rotation is known to place heavy demands on working memory even when mental-state attribution is not involved (Hyun & Luck, 2007). Peter Carruthers has recently argued that this role for mental rotation constitutes a serious confound for many Level-2 perspective-taking tasks, and that these tasks do not so much demonstrate a difference in the concepts of SEEING deployed in Level-1 and Level-2 scenarios or a difference in underlying mindreading systems as a difference in non-

mentalistic task demands (Carruthers, 2015a, 2015c). As an alternative explanation, Carruthers suggests the lack of altercentric interference in the number-perspective task was due to motivational factors: because they were not sufficiently motivated to represent the avatar's perspective, subjects in this task simply did not go to the trouble of mentally rotating the numeral on the table.[21]

One interesting possibility that emerges from Carruthers' motivation-based interpretation is that changing the motivational structure of the number-perspective task could potentially lead participants to maintain a representation of the other agent's Level-2 perspective. Elekes and colleagues investigated this possibility by creating a modified version of the number-perspective task, which subjects either completed by themselves (the Individual condition) or with another participant (the Joint condition) (Elekes et al., 2016). This initial modification of the number-perspective task is especially noteworthy: while a nondescript avatar might be salient enough to warrant Level-1 perspective-taking, it is not obvious that participants would care enough to go to the trouble of maintaining a representation of its Level-2 perspectives. Exchanging the avatar for a live human being both increases the potential salience of the target (real people are generally more interesting than nondescript avatars), and improves the ecological validity of the paradigm. As we'll see shortly, this manipulation proves to be effective.

[21] Carruthers does accept that the evidence from the dot-perspective task shows that Level-1 perspective-taking is automatic, although he denies that these results are best explained in terms of a non-representational concept of seeing. On his "one-system" account, the attribution of mental state concepts is automatic when executive resources are not required, and "spontaneous" when they are. However, the argument from gaze-cueing from the previous section shows that even Level-1 perspective-taking is a spontaneous activity, rather than truly automatic.

Subjects in this experiment completed a number-verification task, which involved rapidly judging whether the number they saw on a screen lying flat in front of them was the same as the number they heard in an audio recording. In the Joint condition, experimenters manipulated whether participants believed that the person seated across from them was completing the same number-verification task (the perspective-dependent task), or an *n*-back task in which subjects had to judge whether or not the color of the number on the screen was the same as the number that came before it (the non-perspective-dependent task). Thus, in both tasks in the Joint condition, subjects knew that their partner was also attending to the numeral on the screen, but only subjects completing the perspective-dependent task believed that their partner was attending to the same aspects of the numeral (namely, its value). But importantly, all subjects ever had to do was complete their own task; their partner's performance was irrelevant.

The experimenters found that subjects in the Joint condition were slower than in the Individual condition, but only when both completed the perspective-dependent task *and* the numerals of the screen were such that their values differed on the basis of perspective (i.e. 2, 5, 6 and 9); for numerals whose values appeared to be the same regardless of which side of the table the participant was at (i.e. 0 and 8), there was no difference between the individual and joint conditions. In effect, subjects were only slower when 1) they had a live partner, 2) they believed that their partner had a similar goal, and 3) the partner's response would diverge from their own on the basis of their Level-2 perspective. These results suggest that knowing that a partner possesses a similar goal to one's own creates an unconscious motivation to maintain a

representation of their perspective, even when this is not relevant to one's overt goal. When this representation differs from one's own first-personal one, this creates altercentric interference.

Using a very similar design, Surtees and colleagues obtained a slightly different set of effects (Surtees, Apperly, & Samson, 2016). Like Elekes et al. (2016), they used a number-verification task that used live partners seated on opposite sides of a display that lay flat on the table between them; in one of the experiments, Surtees et al. also included a version of that task where one partner made judgments about a surface feature of the numeral on the screen, rather than its value. And just like in the Elekes et al. design, subjects only ever had to judge the value of the number from their own perspective – the perspective of the other participant was always task-irrelevant. However, in the Surtees et al. (2016) design, subjects took turns instead of completing the task at the same time; turn-taking either occurred within the same block of trials (with the two participants alternating rapidly), or in separate blocks (with one participant going first and the other going second).

Like Elekes and colleagues, Surtees et al. found that the presence of a live participant affected subjects' Level-2 perspective-taking, with an altercentric interference effect when their perspectives were inconsistent, and also a facilitation effect when their perspectives were the same. But unlike Elekes et al., they found that altercentric interference arose even when the partner was attending to surface features of the numeral, rather than its value. They also found that altercentric inference did not occur in subjects who went first when completing the task in separate blocks;

however, when the second partner took her turn, the altercentric interference effect re-emerged.

Collectively, the results of Elekes et al. (2016) and Surtees et al. (2016) yield a number of conclusions regarding Level-2 perspective-taking, as well as some open questions. First, using a live participant instead of an avatar seems to increase the likelihood that subjects will spontaneously adopt another agent's Level-2 perspective, even when it is not relevant to their overt goals; however, the mere presence of a live participant is not sufficient for this to occur. In the simultaneous task design of Elekes and colleagues, participants only took their partner's perspective into account when explicitly informed that they were performing the same task. In the turn-taking design of Surtees et al., subjects only adopted their partner's perspective when they had previously observed their partner completing the task that they themselves were about to undertake. In both cases, some form of prior knowledge was necessary for spontaneous Level-2 perspective-taking to occur.

The fact that subjects in the Surtees et al. task spontaneously adopted the perspective of their partner even when the partner was not attending to a perspective-dependent feature of the numeral on the screen is inconsistent with the findings of Elekes et al. However, this difference may be due to the difference between the alternating turn-taking task design used in the former study, and the simultaneous task design used in the latter. It is possible that the turn-taking activity created the sense of a shared goal, when in fact there was none.

The most important conclusion to be drawn from this set of findings is that Level-2 perspective-taking can, at times, be fast and efficient, provided that subjects

are provided with the right background knowledge and are sufficiently motivated. This contradicts the claim that Level-2 perspective-taking is a slow and effortful process. In more ecologically valid tasks that use a live participant rather than an avatar, Level-2 perspective-taking turns out to be spontaneous (just like Level-1 perspective-taking).

These findings create something of a puzzle for both the two-systems theorists and its critics, such as Carruthers: if Level-2 perspective-taking tasks place inherent demands on working memory (either because working memory is a constitutive part of explicit mindreading more generally, or because of the mental rotation confound), how come subjects were able to efficiently generate Level-2 perspective representations in these circumstances? The answer may be related to the fact that spontaneous perspective-taking only occurred when subjects possessed the appropriate prior knowledge (in addition to the right motivations). Once subjects learned that their partners' perspective systematically differed from their own (e.g. "If I see 6, he sees 9"), they would have been able to store that knowledge as a mentalistic schema in long-term memory, where it would have been available for rapid retrieval.[22] Thus, even if subjects had to initially engage in effortful mental rotation to judge their partner's perspective, they would subsequently be able to infer their perspective without any effortful spatial reasoning at all. By using memory-

[22] Christensen and Michael (2015) discuss the use of schemas in mindreading at length in their "cooperative multi-systems architecture" proposal, which they offer as an alternative to the two-systems account.

based strategies, subjects would have been able to circumvent the need for any effortful use of working memory.[23]

It is worth noting that Apperly (2010) does discuss one possible way that explicit, demanding forms of mindreading could be rendered fast and efficient: *downwards modularization*. The basic idea behind downwards modularization is that expertise can render otherwise demanding tasks fast and efficient. For example, where an average chess player might discover a path to checkmate through slow, effortful reasoning, an expert player might, thanks to her extensive experience, arrive at a similar conclusion in a seemingly effortless manner. One way that this sort of efficiency-through-expertise can be achieved is when a body of knowledge – initially acquired through explicit, effortful processes – is used so often that it leads to the formulation of cognitive schemas. These schemas enable us to rapidly pair inputs to the appropriate behavioral outputs without having to go through any effortful, explicit reasoning. But, according to proponents of downwards modularization, this efficiency is achieved at the cost of flexibility. Just like innate "original" modules, these acquired modules are ultimately encapsulated from goals and background knowledge. Apperly suggests that downwards modularization might often occur with our explicit mindreading abilities: an expert poker player may, for example, become so well-

[23] Interestingly, Michelon and Zacks discovered that subjects also tended to use memory-based strategies in a Level-1 perspective-taking task: instead of calculating the line-of-sight of an agent directly, participants simply memorized the set of objects that the agent could see, and this led to increased performance (Michelon & Zacks, 2006). The experimenters, who were interested in studying how line-of-sight is calculated, developed a method to control for this strategy. But it highlights the fact that memory-based perspective-taking strategies provide an ever-present, efficient alternative to the use of more spatial forms of reasoning, whether these involve line-of-sight calculation or mental rotation.

practiced that she is able to automatically detect a bluff without needing to engage in any explicit reasoning at all.[24]

However, the effects on Level-2 perspective-taking described above are not plausibly the result of downward modularization. First, subjects never had the explicit goal of monitoring the other agent's perspective at all; Level-2 perspective-taking was actually detrimental to their performance on the explicit task. Expertise, in this context, would consist in ignoring the partner, not representing the way she saw the number. Second, it is implausible that subjects came into the experiment with an acquired module for Level-2 perspective-taking. If this were the case, then altercentric interference should have been present across all the Joint conditions (or, in the case of the Surtees et al. findings, the conditions where partners were merely present, but not yet engaged in the number-verification task), not just the ones where subjects shared a similar goal. The fact that these altercentric interference effects were so context-sensitive suggests that the Level-2 perspective-taking abilities brought by subjects to the lab were flexible and goal-dependent, not stimulus-driven. Thus, the fast and efficient Level-2 perspective-taking that we find in these studies seems to occur in spite of the fact that it is unencapsulated, which runs contrary to the downwards modularization proposal.

6. Implications for the two-systems account

The arguments of the last two sections create serious problems for the two-systems account of perspective-taking. Contrary to that framework, it appears as though both

[24] See Thompson (2014) for a detailed critique of this proposal.

Level-1 and Level-2 perspective-taking can be fast and efficient, but also sensitive to goals and background knowledge. Thus, both forms of perspective-taking appear to occupy the "spontaneous" middle ground between the fast-yet-inflexible and flexible-yet-slow information-processing profiles of the implicit and explicit mindreading systems. In both cases, this combination of flexibility and efficiency seems to be achieved through the interaction between executive systems, long-term memory, and motivational factors. This is not to say that the underlying processes in the two kinds of perspective-taking are really identical: both seem to make use of different cognitive strategies, and are suited to different types of problems. But neither are the two clearly dissociable, as the two-systems framework would suggest.

One obvious conclusion to be drawn from this fact is that there need not be any trade-off between speed and representational flexibility when it comes to our perspective-taking abilities. On its own, this conclusion may not be fatal to the two-systems account: perhaps the distinction between Level-1 and Level-2 perspective-taking does not map onto the implicit and explicit mindreading systems after all, but this framework may still capture important distinctions when it comes to other forms of mental-state attribution. However, the case of perspective-taking should also lead us to view the basic idea of a flexibility-efficiency trade-off in the domain of mindreading with suspicion. Not only does this notion of a trade-off not apply in the case of perception – the domain it was originally intended to explain – but now it has also fallen short in explaining the cognitive underpinnings of one of our core mindreading abilities. Why expect that it should suddenly apply elsewhere?

As a matter of fact, there is evidence that in addition to Level-1 perspective-taking, other forms of "implicit" mindreading also appear to be unencapsulated from background knowledge. For instance, the attribution of motor intentions[25] through motor simulation or "mirror neurons" is often suggested to be automatic and encapsulated from background knowledge. Most commonly, this process is said to involve the automatic mapping of the visual kinematics of an observed action onto the motor system. Our motor system then simulates the performance of that same action, which permits an inference to a guiding motor intention (Rizzolatti & Craighero, 2004) by using our motor planning system in reverse (Jeannerod, Arbib, Rizzolatti, & Sakata, 1995). According to this view, the only inputs to the mirror neuron system are the low-level visual properties of actions.

However, other research on the mirror neuron system is inconsistent with this picture. Monkeys' mirror neurons do not activate for mimicked actions, as when they observe an experimenter pretending to grasp a non-existent object (Gallese & Goldman, 1998); conversely, monkeys' mirror neurons do activate when they witness an occluded grasping action that has no low-level visual properties – but only if they know in advance that there is food behind the occluder (Umiltà et al., 2001). In humans, it's been found that background knowledge about whether or not an observed action is intentional, or whether it has been carried out by an intentional agent, affects the degree to which they are motor-primed to perform that same action (an effect of mirror neuron activity) (Liepelt & Brass, 2010; Liepelt & Cramon,

[25] Motor intentions are intentions to engage in a particular motor action, such as grasping or throwing. These are distinct from distal or future intentions (what I plan to do at some point in the future) and present intentions (what I plan to do now, framed at a level of abstraction that is independent of any particular motor plan) (Pacherie, 2008; Spaulding, 2015).

2008). In other words, the attribution of motor intentions, just like the attribution of Level-1 perspectives, does not seem to be fully automatic or informationally encapsulated. Rather, it is sensitive to background knowledge and abstract features of context. Several authors have taken these findings as evidence that the mirror neuron system actually reflects the effects of a top-down, information-rich form of action prediction, rather than a low-level mapping process (Csibra, 2008; Kilner, Friston, & Frith, 2007).

Further problems for the two-systems account of mindreading arise from studies of "implicit" false-belief[26] tracking (Schneider, Bayliss, Becker, & Dux, 2012). In these tasks, subjects in an eye-tracker passively observe videos of an agent hiding an object and then leaving a room. While the agent is absent, the location of the object is changed. When she returns, subjects look in anticipation towards the previous location of the hidden object (the one last seen by the agent), suggesting that they were tracking her false beliefs. When subjects were debriefed after the task, they showed no sign that they were consciously tracking the agent's belief, suggesting that any belief-tracking that occurred was unconscious and implicit. However, when subjects in the same task are given even a light working-memory task, the implicit belief-tracking effect vanishes (Schneider, Lam, Bayliss, & Dux, 2012). One way of interpreting this finding is to conclude that implicit belief-tracking involves working memory; however, given that the contents of working memory are usually conscious, and subjects reported no conscious belief-tracking, this seems unlikely. What's more

[26] Proponents of the two-systems account would deny that these experiments provide evidence for "belief-tracking," since they hold that the implicit system does not represent "full-blown" propositional attitudes. Rather, they would describe these results as evidence of the tracking of "registrations," a quasi-mentalistic, implicit analogue of beliefs represented by the implicit system (Butterfill & Apperly, 2013).

plausible is that when subjects were engaged in the working memory task, they shifted too much attention away from the agent's perspective for encoding of belief-states to occur or be fixed in long-term memory. Thus, implicit belief-tracking does not seem to be genuinely automatic; rather, as Level-1 perspective-taking, it's likely that we possess a standing disposition to represent the beliefs of others, but only when doing so is either cognitively efficient or somehow goal-relevant.

These findings suggest that other forms of implicit mindreading may also be spontaneous and context-sensitive, rather than automatic and encapsulated. If so, then the entire two-systems framework may be in jeopardy. The principal theoretical motivation for the two-systems account was that fast, efficient, "implicit" processes are likely to be encapsulated, which in turn yields signature limits on their representational capabilities. Instead, we find that implicit mindreading processes are generally quite flexible, and well-integrated with long-term memory, executive systems, and goals. If this is right, then it's not obvious whether there really are any grounds for holding the implicit mindreading system exists.

If the implicit mindreading system is not present in adults, this also casts doubt on the developmental claims of the two-systems view. Part of the two-systems approach to development has been to propose that younger children's early theory-of-mind abilities (e.g. Onishi & Baillargeon, 2005) are products of the implicit mindreading system, and thus subject to signature limits on their representational abilities (Butterfill & Apperly, 2013); in particular, children below the age of four should not be able to pass Level-2 perspective-taking tasks, since these require "full blown" propositional attitude attribution. Proponents of the two-systems account

tested this prediction in two separate studies, and obtained seemingly positive results: infants' looking times did not reflect any Level-2 perspective-taking, and thus seemed subject to signature limits (Low et al., 2014; Low & Watts, 2013). But as with other Level-2 perspective-taking tasks, these paradigms involved mental rotation, and thus potentially confound Level-2 perspective-taking with effortful uses of working memory (Carruthers, 2015c).

When this mental-rotation objection is supplemented by the revelation that the "signature limits" interpretation is based on an erroneous, encapsulated conception of the implicit mindreading system, it becomes all the more clear that these results provide no support for a two-systems account of infant theory-of-mind abilities. If infant mindreading abilities are really subject to any signature limits on their representational capabilities, it is unlikely that these are due to a distinct, encapsulated mindreading system that persists into adulthood. These limitations are more likely the product of immature executive resources, motivational factors, or a lack of relevant experience. Collectively, these factors may create a kind of ersatz encapsulation early in development, but this would dissipate as children's developing executive resources and increasing social experience provides them with a more flexible, integrated set of mindreading abilities.

7. Conclusion: Efficient, context-sensitive mindreaders

Beyond its implications for the two-systems account, this critique highlights some important features of our mature mindreading abilities. One is that several implicit forms of mindreading do not seem to be genuinely automatic; rather, we deploy these capacities selectively, in a context-sensitive, goal-dependent fashion (although we

may also be generally motivated to engage in mentalizing when doing so is cognitively efficient). However, our context-sensitive, goal-dependent mindreading abilities can still be quite fast and efficient. This combination of speed and context-sensitivity seems to be due to the integration of domain-specific mindreading mechanisms with domain-general attentional processes and background knowledge. We also find that even complex, so-called "explicit" forms of mental-state attribution, such as Level-2 perspective-taking, can also be both fast and efficient, provided that we possess the right background knowledge and that we are appropriately motivated.

Another significant conclusion to draw from this discussion is that whether we spontaneously engage in very simple forms of mindreading, or very complex forms of mindreading, or no mindreading at all, seems to be a function of our motivations. Along with our background knowledge, our social attitudes seem to determine the amount of processing resources that go into representing the minds of others. Sometimes, we are highly motivated to represent the mental states of others accurately, and we make use of background knowledge in order to do so quickly and efficiently; at other times, we are less motivated, and as a result our mental state representations are much sparser, as we rely on general-purpose heuristics, such as computing line-of-sight. And, as we saw in many of the gaze-cueing studies, sometimes our background beliefs about the intentional status of an agent or its social group membership give us reason to ignore its perspective altogether. The depth of processing involved in a given mindreading task thus depends on our social goals.

Moreover, as we saw in the discussion of Elekes et al. (2016) and Surtees et al. (2016), the availability of relevant background knowledge enables the mindreading

system with a way to circumvent slower, more effortful forms of reasoning. Notably, in these studies, the relevant background knowledge was not antecedently available to the participants when they first engaged in the task. But when subjects were sufficiently motivated to do so, they were able to generate situation-specific, mentalistic schemas that enabled them to rapidly update their representation of their partner's mental states. In other words, one of the things that we seem to do during social interactions is create shortcuts that make the task of mindreading faster and more efficient – provided, that is, that we are motivated to do so.

Now, contrast this picture of mindreading with the one put forward by mindreading skeptics and endorsed by two-systems theorists (Apperly 2011; Bermudez 2003; Zawidzki 2013). On their view, genuine mental-state attribution consists in a holistic, unbounded form reasoning that parallels the structure of first-person decision-making. According to this picture, mindreaders must, when inferring the mental cause of an action, consider an indefinite range of potential belief-desire combinations. The computational demands of this kind of mental-state inference are surely immense. Clearly, as a theory of how we are able to seamlessly engage in complex forms of coordination or quickly infer intended speaker meanings, this model of mindreading is inadequate; rather, it seems to represent the mental-state attribution strategy of an ideal thinker, unhindered by the demands of computational complexity.

Not being ideal thinkers ourselves, we rarely – if ever – engage in this kind of mindreading. But, contrary to the mindreading skeptic, this does not mean that we rarely engage in mindreading at all. Nor does it mean that we rely on a module for

quasi-mentalistic mindreading, as the two-systems theorists have proposed. Rather, we deploy a range of flexible mentalizing strategies to navigate the social environment, which we tailor to match our situational needs. Some of these strategies may indeed involve effortful, working-memory based forms of cognition. But we do not engage in these effortful reasoning strategies any more than is necessary. Instead, we supplement this kind of reasoning with a number of more efficient strategies. Sometimes, these involve simple, spatial heuristics, as with Level-1 perspective-taking. But we also use more effortful forms of reasoning to create mindreading shortcuts, in the form of mentalistic schemas that may be rapidly retrieved from memory in order to maintain up-to-date models of other people's mental states. And even these more efficient forms of mindreading are deployed in a selective, context-sensitive manner. In short, we economize our mindreading strategies so that they may best fit our needs. We only ever mindread as much as we have to.

Thus, skeptical doubts about the mindreading paradigm can be assuaged once we appreciate the context-sensitive, goal-dependent nature of mental-state attribution. It is a mistake to believe that everyday mindreading consists in a holistic, unbounded form of "central" reasoning. It is also a mistake to argue that if we rarely engage in this idealized form of mindreading, then we must not mindread very much at all. The two-systems view attempted to carve out a middle ground between these two extremes, but it erred in its concession to the skeptic that "full-blown" mindreading must cognitively effortful. With the case of spontaneous perspective-taking, I've shown that our mindreading abilities are much more flexible, efficient and context-sensitive than either the two-systems theorists and the skeptics had thought possible.

Chapter 4: Character and theory of mind: An integrative approach[27]

1. Introduction

As highly social beings, we need to be able to rapidly predict and interpret the behavior of those around us in order to thrive. We do this, the usual explanation goes, by reasoning about the unobserved representational states that cause behavior – a process variously referred to as *theory of mind*, *mindreading*, and *folk psychology*. Standard models of mindreading, such as the theory-theory, the simulation theory, and various hybrid models, tend to focus especially on how we predict and interpret behaviors in terms of beliefs and desires. This focus is epitomized by the field's longstanding fascination with the false-belief task, which is used to measure children's understanding of the representational nature of belief (Onishi & Baillargeon, 2005; Rakoczy, 2015; Wellman et al., 2001; Wimmer & Perner, 1983). As a result, questions about the developmental, cognitive, and evolutionary underpinnings of belief-reasoning tend to dominate social cognition research.

Due to this narrow focus on beliefs and desires, other conceptual tools that we use to interpret behavior are often ignored. One such tool is character-trait attribution: the explanation and prediction of behavior in terms of enduring internal properties of individuals that lead to stable behavioral tendencies. This omission from the theory-of-mind literature is quite curious: one of the most robust findings in social

[27] This chapter was originally published as Westra (2017). It has been reprinted here with permission from Springer, copyright license #4100230216361.

psychology research is that we often interpret behavior on the basis of stable personality traits (D. L. Ames, Fiske, & Todorov, 2011; Gilbert et al., 1995). Character also figures prominently in moral philosophy (Anscombe, 1958; Foot, 1967; Miller, 2013), and has begun to garner attention in empirical moral psychology research as well (Sripada, 2012; Uhlmann, Pizarro, & Diermeier, 2015). Yet in spite of its presence in neighboring disciplines and a large body of data on the subject, character-trait attribution does not figure in classic and contemporary theories of mindreading.

My goal in this paper is to provide a framework for integrating our understanding of character-trait attribution with other aspects of theory of mind. I will propose that we use representations of a person's stable character traits to infer which hypotheses about that person's more transient mental states – namely, their beliefs, goals, and intentions – are more probable; we then use these mental-state hypotheses to directly predict their behavior. Trait attribution thus forms the upper level of an action-prediction hierarchy, wherein the hypotheses at higher levels inform the hypotheses at lower levels. Feedback from observable behavior then leads us to make revisions to our mentalistic hypotheses, which might occur at either the belief-desire levels or at the level of character traits. This basic inferential structure is best understood in terms of a hierarchical Bayesian model of cognition.

In section 2, I will briefly review part of the empirical literature on the attribution of character traits, and the role that these representations play in predicting and interpreting behavior. In section 3, I will discuss recent "pluralist" accounts of

folk psychological reasoning (Andrews, 2008, 2012; Fiebich & Coltheart, 2015), which *do* acknowledge the role of character-trait attribution in folk psychology, but fail to explain its relationship to other forms of mindreading. In sections 4 and 5, I will outline an account in which character-trait attribution stands in a systematic, hierarchically structured relation to belief and desire attribution. In section 6, I suggest several ways to empirically test this account, as well as ways to apply it to other, related domains.

First, however, a word about how we think of character traits. Like beliefs and desires, character traits are believed to be causally related to behavior, and it is not uncommon to explain behavior by referring to a character trait (e.g. "She turned in the lost wallet because she is an honest person") (Malle, 2004). Some traits, such as selfishness and greed, possess a strong volitional element, and thus seem closely related to desires. Others, such as intelligence, cleverness, and gullibility, seem distinctively epistemic, and thus more related to beliefs. However, traits also differ from beliefs and desires in several important ways. First, beliefs and desires figure in practical reasoning, and lead directly to action. Character traits, on the other hand, do not seem to figure in practical reasoning, and it is less clear how they translate into particular actions. Second, beliefs and desires can easily change. When we acquire new relevant information we regularly change our beliefs; when we successfully act out our plans, our desires and goals are fulfilled. Character traits, conversely, are much more persistent. They do not update or go away as the result of individual actions, but rather last through significant portions of an agent's lifetime. Third, beliefs and desires can be about particular states of affairs. Character traits, on the

other hand, relate to the world in a very general way, and tend to be relevant across a wide range of situations. In short, character traits are temporally stable mental properties that relate to action in an opaque, general manner across a wide range of situations.[28]

Citing evidence from social psychology, some philosophers have questioned whether people actually possess character traits as we ordinarily think of them (Doris, 2002; Harman, 1999). These arguments begin with findings showing that subtle manipulations in situational factors lead to dramatic effects on behavior. For instance, Isen and Levin showed that finding a dime in a phone booth makes people much more likely to help a stranger pick up dropped papers – a finding that seems to show that "generosity" is, contrary to common belief, a fickle, variable trait (Isen & Levin, 1972). Based on these and other experimental results, "situationists" have argued that it is situations, and not stable character traits, that really cause our behavior. These arguments have sparked a great deal of controversy, and a number of philosophers have mounted defenses of the reality of character traits(Miller, 2013; Sabini & Silver, 2005; Sreenivasan, 2002).

The situationism debate is about the metaphysical reality of character traits, and whether stable character traits ought to figure in mature scientific explanations of

[28] John Doris offers a similar, though not identical, analysis of character traits. According to his view, "global" traits have two primary features:

1. *Consistency*. Character and personality traits are reliably manifested in trait-relevant behavior across a diversity of trait-relevant eliciting conditions that may vary widely in their conduciveness to the manifestations of the trait in question.
2. *Stability*. Character and personality traits are reliably manifested in trait-relevant behavior over iterated trials of similar trait-relevant eliciting conditions. (Doris, 2002, p. 22)

Doris also mentions a third feature of character traits, *evaluative integration*, which is not relevant for our current purposes.

behavior; situationists think that insofar as character traits exist, they are situation-dependent, and that stable character traits have no real explanatory value for psychology. This is not a paper about the metaphysical reality of character traits, however. Rather, it is about people's *representations* of character traits, and the role that these representations play in *folk-psychological inference*. Notoriously, folk reasoning about the world is often inaccurate, and frequently invokes entities that do not stand up to scientific scrutiny. For instance, when reasoning about the motion of objects, even educated adults seem to rely on a quasi-medieval impetus principle, and predict that (for example) a ball spinning around the end of a string will continue to follow a curved path when it is released (McCloskey, Caramazza, & Green, 1980). Impetus principles no longer play any role in our mature physics, but they still play a role in folk physics. Likewise, stable character traits might have no place in our mature psychology, but they clearly still play a role in our everyday social reasoning. Thus, the situationists might be right that our behavior is never caused by stable character traits, but they could still allow that representations of stable character traits play an important role in our folk psychology. A dedicated situationist could thus happily accept the foregoing account as an error theory explaining why we think people have stable character traits, even though there are none. The two views are, in principle, mutually compatible.

However, while the present account is not committed to the existence of stable character traits as such, it does imply that our representations of character traits have *some* predictive value; otherwise, their prominent role in our cognitive economy would be mysterious. The most obvious explanation for this predictive role would be

that stable personality traits do in fact exist, despite the evidence cited by the situationist. Another explanation would be that trait representations serve as a kind of inferential heuristic: even if they fail to track anything real in the world, they may earn their predictive keep by conferring some sort of information-processing benefit on other socio-cognitive processes (such as belief-desire reasoning). Along these lines, I suggest that one function of trait attributions is to provide us with initial prior probability distributions over mentalistic hypotheses, which then undergo further updating in response to experience. Another explanation would be that trait representations roughly track *something* in the world – just not character traits. Like the impetus principle, which roughly tracks the real physical principle of inertia (but systematically errs in certain cases), it may be that our trait representations roughly correspond to some predictively relevant property of the social environment, which we *systematically misrepresent* as stable character-traits. In this vein, I suggest in section 6 that our trait attributions may sometimes track relational social properties such as status and intergroup threat.

In order for my account to be right, at least one of these explanations must be correct. It could be that all of them are right: perhaps our trait attributions sometimes track real personality traits, while also tracking relational social properties, while simultaneously conferring an information-processing benefit on our overall action-predictions. However, I need not take a strong stand on this issue at this time. All that matters for my current purposes is the role that trait information plays in the structure of mentalistic action-prediction.

2. Impression formation and mindreading

In social psychology and socio-cognitive neuroscience, reasoning about character traits is most often referred to as 'impression formation' and 'person perception' (D. L. Ames et al., 2011; Trope & Gaunt, 2007). While we attribute a wide range of specific traits to others, it appears that the kinds of traits we appeal to tend to be organized along two particular social dimensions: warmth and competence (Fiske, Cuddy, & Glick, 2007). The warmth dimension captures attributions of traits such as friendliness, sincerity, trustworthiness, and seems to track whether we expect an individual to be positively or negatively disposed towards us. The competence dimension, in contrast, captures attributions of traits such as intelligence, impulsivity, and social dominance, and seems to track our estimation of an individual's ability to successfully achieve their goals. When trait attributions are analyzed in terms of these two dimensions, they are highly predictive of our reactive attitudes towards both individuals and groups (Cuddy, Fiske, & Glick, 2007), even across a wide range of cultures (Cuddy et al., 2009).

Interestingly, these two trait dimensions do not just emerge in people's judgments of individuals: they also emerge in stereotypes about social groups. For instance, common anti-Semitic stereotypes tend to invoke low warmth traits, such as deceptiveness and miserliness, but also high competence traits, such as intelligence. Misogynistic stereotypes, in contrast, invoke high warmth traits, such as helpfulness, but also low-competence traits, such as frivolity and superficiality. Stereotypes about very low status groups, such as the homeless, tend to contain low competence traits,

such as stupidity and laziness, and low warmth traits, such as dishonesty. Both social in-groups (e.g. students if one is a student) and societal prototype groups (in Western cultures, the White middle class) tend to be rated as both high competence and high warmth (Cuddy et al., 2009; Fiske et al., 2007; Fiske, 2015). This suggests that we may use a person's social group membership as a source of evidence about her character traits (see also Fiebich and Coltheart (2015)).

Many of the methods used to study trait attributions involve explicit, linguistically based measures (e.g. Ross 1977); however, character-trait attributions can also be extremely rapid and unconscious. In particular, we seem to use low-level perceptual cues such as facial appearance to make character-trait attributions within 100 milliseconds of encountering someone (Moshe Bar, Neta, & Linz, 2006; Todorov, 2013). Incredibly, these rapid, perceptually based trait attributions also vary along the dimensions of warmth and competence (or, as they are referred to in this literature, trustworthiness and dominance). In particular, neutral-expression faces with wider jaws, heavier brows, and smaller eyes tend to be judged as more dominant, while "baby faces" tend to be judged as less dominant; similarly, neutral-expression faces with downturned brows and lips tend to be judged as less trustworthy, while faces with high brows and upturned lips tend to be judged as more trustworthy (Todorov, Said, Engell, & Oosterhof, 2008). Thus, from the first second that we encounter someone, we begin to form a representation of his or her character traits along the warmth and competence dimensions.

Of course, we do not infer personality traits from appearance alone: we also use a person's behavior and second-hand information to inform our representations of their character. The most well-known finding in this vein is that we tend to over-attribute the causes of behavior to underlying traits or dispositions rather than situational factors – a phenomenon known as the 'correspondence bias' or the 'fundamental attribution error' (Gawronski, 2004; Gilbert et al., 1995; E. Jones & Harris, 1967; Ross, 1977). For instance, when participants passively observe one confederate quizzing another, they tend to rate the questioner as more intelligent than the test-taker, even though the questioner has clearly been provided with the answers, while the test-taker has not (Ross 1977). This bias can be mitigated by prompting participants to explicitly attend to situational factors (e.g. that the questioner has been provided with all the answers, whereas the test-taker has not); however, participants will default to the correspondence bias when placed under cognitive load, even if the very same situational information is made explicitly available (Gilbert et al., 1995; Gilbert, Pelham, & Krull, 1988). This suggests that the correspondence bias is the product of a relatively efficient cognitive process, while correcting it requires cognitive control.[29]

[29] There are also cross-cultural differences in the extent to which individuals fall prey to the correspondence bias. While the correspondence bias is present to some extent across cultures (Choi, Nisbett, & Norenzayan, 1999; Krull et al., 1999; Norenzayan, Choi, & Nisbett, 2003), it appears that members of "individualist" societies are particularly susceptible to it; meanwhile, members of "collectivist" societies seem to pay more attention to situational factors and the presence of social constraints (Choi & Nisbett, 1998; Miyamoto & Kitayama, 2002). This is consistent with the broad finding that members of collectivist cultures display a habitual tendency to attend to situational factors and contexts (Kitayama, Duffy, Kawamura, & Larsen, 2003). These habitual patterns of attention seem to make members of "collectivist" cultures better able to correct their initial dispositionalist attributions.

Much of the research on the correspondence bias has occurred separately from research on mindreading, focusing instead on the distinction between situation-based and disposition-based explanations of behavior. But some social psychologists have begun to explore the connection between representations of traits and other mental states, such as intentions. For instance, Krull and colleagues found that participants were less likely to exhibit a correspondence bias when an actor performed a helpful action if the actor showed signs of unwillingness (Krull, Seger, & Silvera, 2008). Likewise, a number of authors have found that the correspondence bias was attenuated when participants were given reason to think that a given action may have been performed for an ulterior motive (D. R. Ames, Flynn, & Weber, 2004; Fein, 1996; Reeder, Vonk, Ronk, Ham, & Lawrence, 2004). Hooper and colleagues also found that participants who were primed to engage in explicit perspective-taking displayed a diminished correspondence bias compared to a control group (N. Hooper, Ergogan, Keen, Lawton, & McHugh, 2015). Thus, thinking about mental states seems to mediate inferences from behavior to character (Reeder, 2009).

This relationship between trait attribution and other forms of mental-state attribution is reflected in the neural correlates of both processes, which overlap substantially and appear to be functionally related. Many neuroimaging studies have confirmed the existence of a distinctive network of brain regions that are consistently recruited when we reason about the thoughts and behavior of others: the temporal-parietal junction (TPJ), the posterior superior temporal sulcus (pSTS), the medial prefrontal cortex (mPFC), the precuneus (PC), and the temporal poles (TP) (Van Overwalle, 2009). All of these regions are implicated in impression formation and

updating, both under intentional and spontaneous conditions (Cloutier, Gabrieli, O'Young, & Ambady, 2011; Ferrari, Vecchi, Todorov, & Cattaneo, 2016; L. T. Harris, Todorov, & Fiske, 2005; Hassabis et al., 2013; Kestemont, Vandekerckhove, Ma, Van Hoeck, & Van Overwalle, 2013; Ma et al., 2011). The dorsal region of the mPFC (dmPFC) in particular seems to be centrally implicated in the representation of stable personality traits (Ferrari et al., 2016; Ma, Vandekerckhove, Van Hoeck, & Van Overwalle, 2012). When subjects are explicitly prompted to reason about traits, this region is highly active; in contrast, when subjects are prompted to reason about "situational" factors, this region is less active, while regions associated with goal and belief attribution, such as the TPJ and the pSTS, are more so (Kestemont et al., 2013). However, when subjects learn that a person holds a belief or performs an intentional action that is inconsistent with a previously formed impression, both the TPJ and the dmPFC show increased activity (Cloutier et al., 2011; Ma et al., 2011). Thus, the neuroimaging data, like the behavioral data, suggest that mental state information interacts with the trait attribution process.

There is also some indication that representations of character traits can bias our mental-state attributions. This evidence comes from a debate about how to interpret the side-effect effect, which is when participants seem to over-attribute intentionality and blame to agents whose actions have negative (but not positive) side-effects (Knobe, 2003). Chandra Sripada has suggested that this effect may be driven by an initial negative judgment of the agent's character, or "Deep Self" (Sripada, 2009). According to this view, participants incorrectly judge that the agent intentionally caused a particular outcome because this intentionality attribution seems

to follow from their previous impression of the agent's character; in other words, they interpret the agent's actions (and their consequences) as flowing from their deeper personality traits. To test this theory, Sripada asked participants who had completed a side-effect effect task to give an explicit evaluation of the agent's character and core values; sure enough, these predicted their intentionality judgments (Sripada & Konrath, 2011; Sripada, 2012).

To summarize: upon first encountering an individual, we rapidly construct a representation of their character that is especially sensitive to particular trait-dimensions, namely warmth and competence. We use various sources of information to update this representation, and are biased towards interpreting behavior as reflecting stable character traits. However, these inferences are mediated by inferences about mental states: when information about the motivations and beliefs of others is available, we update or refrain from updating our character models accordingly. Moreover, background knowledge about character also seems to affect our intentionality attributions, suggesting that we expect not just behavior, but also intentions to accord with character. Inferences about character and inferences about mental states, in short, appear to inform one another.

3. Character-trait attribution and theories of folk psychology

Traditional accounts of mindreading, such as the simulation theory and the theory-theory, have not typically addressed how we attribute character traits. The notion of character seems particularly hard to integrate into a simulation-based account. According to the simulation theory, if an interpreter has enough information about

another agent's beliefs and desires, she can make a successful behavioral prediction by simulating in an offline manner how she would behave if she had those same beliefs and desires (Goldman, 2006; Heal, 1996). But it is not at all clear how this strategy could extend to trait attribution. Character traits are not the sorts of things that could figure in practical reasoning.[30] 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. Thus, it is not clear where – if anywhere – character traits could fit into a pure simulationist account.

The only plausible option for the simulation theorist would be to endorse a hybrid account. Instead of holding that character enters into the simulation process itself, a hybrid simulation/theory-theorist could hold that character information is used to infer the *inputs* to the simulation procedure. This solves the simulationist's character problem, but only at the cost of conceding that simulation theory is poorly equipped for reasoning about traits. It also raises a new question: how would a theory-theorist explain the effects of character on practical reasoning?

Fortunately, the theory-theory seems better equipped to deal with trait attribution. Traditional theory-theory accounts focus on generalizations about how beliefs and desires combine to produce behavior, treating both as underlying causal variables (Gopnik & Wellman, 2012; Wellman, 2014). Quite conceivably, character

[30] Beliefs about one's own character traits could figure in practical reasoning (e.g. "I know that I am an impetuous person, so I should be careful not to act without thinking"). But this observation is of little help to the simulation theorist: surely, this kind of self-reflection is uncommon in the first-person case, and it would be bizarre if we nevertheless believed that other people frequently engage in it. Moreover, beliefs about one's character seem like they would have a very different effect on behavior than character itself. If I reflect on my own impulsivity, for instance, it will probably lead me to be *less* impulsive.

traits could be treated as another kind of underlying variable, albeit one that has a much less direct effect on behavior than beliefs and desires. But any such account would need to do more than just posit an additional variable: it would also have to tell us just how traits relate to mental states, and how they help us to predict behavior. The account that I propose in this paper, which could be construed as a version of the theory-theory, will attempt to do just this.

There is, however, one group of mindreading theorists that has already explicitly addressed the role of character-trait attribution in action prediction and interpretation: the folk-psychology pluralists (Andrews, 2008, 2012; Fiebich & Coltheart, 2015). The basic goal of the pluralists is to offer an alternative to simulation theory and theory-theory accounts that invokes a broader set of procedures and representations. Andrews (2008, 2012) suggests that in addition to belief-desire reasoning, we also use social norms, stereotypes, situation-based schemas, and trait attribution to predict and explain behavior. Likewise, Fiebich and Coltheart (2015) propose that we use trait-based inferences to predict and interpret behavior, in addition to theory-based and simulation-based procedures. They also situate these various socio-cognitive strategies within a two-systems framework[31] (Apperly & Butterfill, 2009; Kahneman, 2011). Thus, when predicting and interpreting another agent's behavior, we may use either System 1 or System 2 versions of simulation, theory, or trait attribution.

[31] System 1 strategies are "fast, relatively effortless routines that occur without our awareness," while System 2 strategies are "slow routines which require the expenditure of mental effort and are subject to consciousness and deliberative control" (Fiebich & Coltheart, 2015, p. 238).

While they differ in several respects, these pluralist accounts treat reasoning about behavior in terms of character traits as a socio-cognitive *alternative* to belief-desire predictions and explanations. Even if an agent did not possess the concepts of BELIEF and DESIRE, according to the pluralists, they might still successfully predict behavior by using their knowledge of a target's traits. This is possible, the pluralists argue, because trait-based interpretations rely on associations between behaviors and situations. For instance, a trait like generosity might form the central node[32] in a network of behavior-situation pairings: *leaving large tips* and *restaurants*, *carrying heavy boxes* and *friends moving house*, and so on. These associative networks would enable agents to attribute traits whenever an individual demonstrated one of the relevant behavior-situation pairings, and then use this information to predict that individual's behavior in other situations in which generosity might be possible.

While the pluralists should be given due credit for emphasizing a role for traits in our folk psychology, this particular approach to trait attribution has two important limitations. The first is that the predictive utility of trait-based reasoning will depend heavily on how 'situations' get represented. If trait-behavior associations

[32] Fiebich & Coltheart distinguish between non-linguistic trait attributions and linguistic trait attributions. Non-linguistic trait attributions occur when an agent does not possess a linguistic concept of a trait (i.e. the word 'generosity'). These only consist in associations between particular behaviors, situations, and agents, and would only allow for predicting similar behaviors in similar situations. Linguistic trait attributions, in contrast, involve the possession of a linguistic concept of a trait, and would facilitate a whole network of predictions.

I am skeptical of this distinction for two reasons. First, non-linguistic trait attribution, on this account, does not seem to involve trait-based reasoning at all: traits are supposed to be enduring, internal properties of individuals, but these non-linguistic trait attributions seem to consist only in superficial behavioral associations. Second, this distinction implicitly assumes that the only way to possess a concept of a trait is through language. But there is ample reason to think that even pre-linguistic or non-linguistic entities can possess concepts (e.g. Call and Tomasello 2008; Carey 2009). While linguistic concepts undoubtedly enrich and expand our trait attribution abilities, there is no reason to think that non-linguistic trait attribution is as impoverished as Fiebich and Coltheart (2015) make it out to be.

are tied only to situations that we have previously experienced, then it will be inert whenever we encounter a novel situation. Andrews (2008) recognizes this fact, but suggests that it is not a big problem, because we spend most of our time in relatively familiar situations. But this reply raises an important question: how do we parse situations for the purpose of trait-based predictions? If we parse situations at a fairly coarse level, then Andrews might be right; however, this would make the corresponding predictions far less reliable, since they would be insensitive to important situational differences. For instance, if one forms the association between *leaving large tips* and *restaurants*, then we would predict that a generous person would leave a large tip even when she has received poor service, or when a friend is treating her. Conversely, if situations are parsed very finely, then we will treat otherwise familiar situations as novel, given a very slight change. Thus, we might expect a generous person to leave large tips in *sushi restaurants with good lighting*, but not in *sushi restaurants with bad lighting*. In this case, most trait attributions would be predictively inert. Unless we parse situations just right, in other words, the pluralist strategy will either lead to inaccurate overgeneralizations, or inflexible under-generalizations.

Some pluralists may simply wish to concede the limited reliability of trait attributions. Andrews (2008) suggests that when we make inaccurate predictions, we may simply respond by forgetting them, or by giving a post-hoc explanation of our failures, and then carry on with our business. This is not a problem, the pluralist argues, because we have lots of different strategies for folk-psychological prediction: when trait attribution fails us, we may simply try a different one. This response is

unsatisfying: even if trait-based reasoning is often inaccurate, it seems that one should be able to learn from these inaccuracies in order to inform future predictions, rather than simply discard them. Indeed, the evidence reviewed in the previous section indicates that we actually pay close attention when our trait-based expectations are violated. But if traits were truly an unreliable way to predict behavior, then why would we continue to track character information? If the pluralist story about trait attribution is correct, then this seems like a bizarre use of limited cognitive resources.

The second, related limitation of the pluralist account of character traits is that it cannot explain the empirical relation between trait attribution and mental-state attribution. As we saw in the previous section, 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. Indeed, Andrews (2008) argues that there is really a double dissociation between belief-attribution and trait-based reasoning. First, while children are able to reason explicitly about beliefs from an early age, they do not explicitly mention traits in their explanations and predictions of behavior until much later (Kalish, 2002). Thus, it is possible to reason about mental states even if one cannot reason about character traits. Second, interventions for children with autism (who lack the ability to reason about beliefs) often rely upon training children to associate traits and behaviors, such as the term ‘happy’ with smiling and laughing (Gray, 2007). Thus, one can also reason about traits without being able to reason about mental states.

There are a few problems with this argument. First, there is now positive evidence that three to four year-old children respond in an adult-like manner to facial features associated with warmth and competence, despite the fact that they do not refer to such traits in their explanations of behavior (Cogsdill, Todorov, Spelke, & Banaji, 2014). Second, the autism intervention Andrews describes does not seem to be about trait-reasoning at all, but rather reasoning about emotions. But even if it were an instance of trait-reasoning, this would then be an exception that proves the rule: in the absence of the capacity to reason about beliefs, it seems that children with autism are only able to use trait information through explicit, laborious training, whereas it comes naturally to neurotypical individuals.

However, even if we were to accept Andrews' double dissociation argument in its entirety, all it would show is that character reasoning and belief-desire reasoning are not *identical*, and that neither is *necessary* for the other. But these are only the strongest possible relations that could hold between these two processes. Even if, in principle, there were double dissociations between character reasoning and belief-desire reasoning, it might still be true that the two processes are causally and functionally intertwined.

I suggest that the solution to the pluralist's first problem may lie in developing an answer to the second. What the pluralist proposal lacks is a principled basis for parsing situations for the purpose of behavioral prediction. But if we consider an agent's beliefs and desires, the solution to this problem is obvious: the 'situation' will consist in those features of the local context that the agent believes are relevant to her

goals. Moreover, this approach would facilitate predictions even in highly unfamiliar situations. This is because mental-state reasoning is a highly flexible, generative framework for predicting and interpreting behavior. By employing causal models that treat mental states as variables that can take on a broad range of values, mentalistic reasoning is capable of generating behavioral predictions about an indefinitely large number of novel situations, even if they have yet to be encountered (Christensen & Michael, 2015). Thus, the predictive link between traits and behaviors postulated by pluralists only makes sense when it is mediated by belief-desire reasoning, because belief-desire reasoning provides us with a principled basis for parsing situations. However, this leaves us with the same lingering question that we started with: how are trait representations related to representations of beliefs and desires?

4. The action-prediction hierarchy

In this section, I introduce hierarchical predictive coding accounts of cognition, and how they have been applied to theory-of-mind research. This will lay the groundwork for my positive account of how representations of character relate to belief-desire reasoning. I begin with a brief digression about the nature of mirror neurons. The purpose of this digression will be to illustrate how predictive-coding approaches are poised to explain the cognitive underpinnings of action prediction. In particular, these accounts posit that we represent intentional states in a hierarchical fashion, and that mirror-neuron activity reflects the way we exploit this hierarchy when predicting intentional actions. Ultimately, I will argue that our representations of character are a part of this action-prediction hierarchy.

4.1. Mirror neurons and action hierarchies

Mental states vary with respect to their temporal stability. For instance, some desire-like states, such as motor intentions – the intention to make a particular bodily movement, such as reaching or grasping – are highly transient. We also chain together many individual motor intentions in order in order to fulfill particular action-goals, as when we walk across a room to pick up a tool; these goals last longer than the individual motor intentions that comprise them, but are still extinguished relatively quickly. Many of these action-goals can be chained together to achieve more complex, temporally extended goals, such as building a house or fixing a car. These broader goals can in turn serve as sub-goals for still larger projects, and so on. Desire-like states, in other words, seem to form temporal hierarchies: more stable goals are comprised of more transient sub-goals, which are comprised of even more transient sub-sub-goals, eventually bottoming out in very low-level motor plans.

This property of desire-like states has not gone unnoticed by mindreading theorists. In particular, it has caught the attention of several authors who were unsatisfied with the standard, "direct-mapping" interpretation of mirror-neuron activity endorsed by simulation theorists (Gallese & Goldman, 1998). According to this standard interpretation, when we observe the low-level visual properties of another agent's movements, we automatically form an offline representation of those actions in the pre-motor cortex, where we normally represent our own action plans. Based on this representation, the story goes, we are then able to deduce the higher-order intentions that would have caused this action plan, and thereby infer the agent's

goals, effectively using our own motor planning system in reverse (Jeannerod et al., 1995; Rizzolatti & Craighero, 2004).

There is a big problem with this account: the inference from low-level visual properties of behavior to goals is vastly under-determined. This is because a single behavior is, in principle, compatible with a wide range of underlying motivations. One could, for instance, raise one's hand with an open palm because one is about to salute, to give a high five, to wave in greeting, to tell someone to stop, or to deliver a slap. The same behavioral effect, in other words, could have indefinitely many different mental causes (a predicament known as an 'inverse problem') (Csibra, 2008; Jacob & Jeannerod, 2005; Kilner et al., 2007). This means that for any given behavior that the mirror neuron system represents, we must sort through an indeterminately large space of possible goals that might have caused it. Thus, the direct-mapping account quickly leads to computational intractability.

This observation has led several authors to argue that mirror-neuron activity does not reflect a bottom-up mapping from motor-intentions to goals, but rather a top-down prediction about an agent's likely behaviors based on a prior hypothesis about its goals – what Csibra (2008) calls the 'predictive action monitoring hypothesis'. This solves the aforementioned under-determination problem, because all it involves is *checking* whether an observed behavior would be made likely given a hypothesized goal, rather than *solving* for a unique goal from an ambiguous behavior. And since goals are a more abstract kind of representation than motor intentions, they tend to be consistent with a fairly wide range of more concrete physical behaviors. For instance,

the goal of eating an apple makes a number of behaviors more likely: reaching over to grab the apple and bringing it to one's mouth, or reaching over to grab an apple and then grabbing a knife to peel it, or using the knife to cut it into wedges, etc. Our action-prediction system solves this problem by selecting the behaviors that are most probable given the goal in question. The computational dilemma faced by the direct-mapping account is thus avoided by taking a top-down, predictive approach that exploits the hierarchical structure of intentional action.

The predictive action-monitoring hypothesis also seems to be more consistent with the existing mirror-neuron data: Gallese and Goldman found that monkeys' mirror neurons show no activity for mimicked actions, as when an experimenter pretends to grasp a non-existent object (Gallese & Goldman, 1998). Umiltà and colleagues also found that monkeys' mirror neurons do respond to actions where low-level visual input is unavailable, as when they watch an experimenter reach behind an occluder to grasp a hidden piece of food (Umiltà et al., 2001). In humans, motor-priming (an effect of mirror neuron activity) seems to be sensitive to background knowledge about the intentional status of an bodily movement: if participants believe that a movement is forced, rather than goal-directed, no motor priming occurs (Liepelt & Brass, 2010; Liepelt & Cramon, 2008). Thus it seems that mirror neurons are responsive to *expectations* about goal-directed action, rather than the low-level visual properties of action.

4.2. Hierarchical Predictive Coding and theory of mind

The predictive action-monitoring hypothesis reflects a growing trend in cognitive science and neuroscience towards predictive models of cognition (Clark, 2015; Seligman, Railton, Baumeister, & Sripada, 2013). The brain, in a very general sense, needs to be in the business of making predictions about the world: without being able to predict what's coming next, planning one's future actions is impossible. These predictions need to happen at multiple timescales simultaneously, whether we are predicting the objects in the space before us as we move through it, predicting where to find food in our local environments, or predicting events in the distant future. Predicting the behavior of other agents, in this sense, is just one part of the larger cognitive challenge of planning one's actions. As creatures that engage in complex forms of social coordination, this kind of prediction is especially important for human beings. The predictive action-monitoring hypothesis thus accounts for a key aspect how human beings are able to form plans at multiple timescales in a highly social environment.

Hierarchical predictive coding theories (HPC) have proven a fruitful way to translate this broad insight about the importance of prediction in cognition to specific hypotheses about neural processing. According to HPC theorists, our expectations about the environment begin on the shortest possible timescale, with predictions about the causes of our present sensory experiences. On this view, neural systems do not just respond to incoming environmental information in a bottom-up manner, but also make forward-looking predictions about what that information will be, which they pass down the cortical hierarchy to the relevant input systems. Incoming

information is then checked against the prediction signal; if the two do not match, an error signal is propagated back up the hierarchy, and checked against the higher order prediction. If these error signals are large, then the information they carry is incorporated into a revised internal model of the causal structure of the world, which then generates new predictions about incoming information. This process then repeats itself iteratively until prediction error signals are minimized. Formally, this account is said to be equivalent to a Bayesian updating procedure, wherein the posterior probability of a given hypothesis is a function of the prior probability of that hypothesis and the probability of a given observation (Moshe Bar, 2007; Clark, 2015; Friston & Kiebel, 2009; Hohwy, 2013; Spratling, 2008).[33]

Building on initial applications of the HPC framework to mirror neurons (Kilner et al., 2007), a number of authors have now proposed HPC accounts of mindreading (de Bruin & Strijbos, 2015; Hohwy & Palmer, 2014; Koster-Hale & Saxe, 2013). The most detailed of these proposals to date is that of Koster-Hale and Saxe (2013). Reviewing a wide range of neuroscientific evidence, they argue that much of the neural activity during mindreading tasks displays the signature of a predictive-coding architecture – namely, greater responsivity to unexpected stimuli than expected stimuli (i.e. prediction error signals). For example, they describe how the STS, which is associated with the processing of biological motion and goal-directed action, displays enhanced responses to unexpected behaviors, either because

[33] There is considerable variation amongst the different versions of predictive coding. Some theorists have taken the extreme position that prediction error signals are the *only* information carried via bottom-up input systems (Clark, 2015; Friston & Kiebel, 2009; Hohwy, 2013), while others allow that traditional bottom-up information-processing compliments top-down prediction (Moshe Bar, 2007; Spratling, 2016). On my account, trait information (e.g. via facial features) is sometimes initially processed in a bottom-up fashion; as such, I disavow the idea that bottom-up input systems only carry prediction errors.

they are inefficient (Brass, Schmitt, Spengler, & Gergely, 2007; Deen & Saxe, 2012) or inconsistent with previously displayed desires (Jastorff, Clavagnier, Gergely, & Orban, 2011). Likewise, the TPJ, which is known to respond to information about beliefs (Saxe & Kanwisher, 2003), displays a stronger response to belief ascriptions that are surprising than those that are expected, given one's background beliefs about an individual (Cloutier et al., 2011; Saxe & Wexler, 2005).

Koster-Hale and Saxe also argue that the data on trait-sensitive activity in the dmPFC is also consistent with a prediction-error minimization account. For instance, after Ma and colleagues provided participants with verbal information about the behavior of an individual (from which various character traits could be inferred), they presented them with test sentences that were either consistent or inconsistent with these descriptions (e.g. "Tolvan gave her brother a *hug*" versus "Tolvan gave her brother a *slap*"). They saw increased responsivity in the dmPFC for trait-inconsistent behaviors (Ma et al., 2011; see also Behrens, Hunt, & Rushworth, 2009; Kestemont et al., 2013; Mende-Siedlecki, Cai, & Todorov, 2013). Thus, the dmPFC seems to be sensitive to prediction errors related to personality traits.

Currently, the HPC approach to theory of mind is still in its infancy. As Koster-Hale and Saxe note, more empirical work needs to be done to develop the positive predictions of this kind of account in detail. However, HPC gives mindreading theorists a well-supported, general empirical framework for explaining the nature of action-prediction that has already been fruitfully applied to a number of different cognitive domains. It also coheres with a broader consensus among

cognitive scientists about the centrality of predictive processes in cognitive systems. Although it is not without its controversies[34], the HPC approach in general is currently a progressive scientific research program (Lakatos, 1970), and a promising way to pursue questions about the nature of social cognition. In the next section, I use this approach to develop an empirically supported conjecture about the relationship between character-trait attribution and other forms of mindreading.

5. Character and the action-prediction hierarchy

Within a hierarchical Bayesian framework, a possible relationship between character traits and other mental states starts to emerge. As we saw initially with the case of mirror neurons, predictions about more transient states of affairs, such as motor intentions, tend to be informed by hypotheses about more temporally stable goal states. Hypotheses about goals, in turn, are informed by representations of more enduring desires. At each subordinate level in the predictive hierarchy, expectations about more transient states are shaped by superordinate hypotheses about more enduring states. As I discussed in the introduction, a key feature of character traits that distinguishes them from beliefs and desires is their greater temporal stability. As such, traits seem to fit naturally into the upper levels of the hierarchy for action-prediction. Background beliefs about character traits could thus inform and constrain

[34] For example, the explanatory status of the Bayesian aspect of these models is a vexed question. Some theorists are explicit that the Bayesian formalism is intended to capture only the computational level of description, abstracted away from implementational, mechanistic details (Chater, Tenenbaum, & Yuille, 2006), while others seem to be making claims about the actual algorithms that support predictive processes (Friston & Kiebel, 2009). While some have charged that ultimately, Bayesian models amount to "just-so" stories with little explanatory value (M. Jones & Love, 2011), there are reasonable answers to such challenges (Zednik & Jäkel, 2016), and plausible ways to interpret the various aspects of Bayesian models that render them empirically tractable (Icard, 2016).

predictions about more transient mental states, which then inform predictions about observable behavior.

To illustrate: suppose that you are observing Tom, whom you believe to be dishonest. A woman walks past, and accidentally drops her wallet in front of him. Tom looks toward the wallet, and then looks back at the woman. Because you know him to be dishonest, you assign a high probability to the hypothesis that Tom desires to steal the wallet. Given this desire-attribution, you might then expect that Tom will perform a series of actions: look around to see if anyone is watching, bend over discretely by the wallet as if tying his shoe, pick up the wallet and put it in his pocket. The prior trait attribution – dishonesty – thus serves as an overhypothesis, raising the prior probability of mental-state hypotheses that are consistent with the trait in question – namely, self-interested desires (Kemp et al., 2007).[35] Thus, when we observe Tom's actions in a particular scenario, the first desire-hypotheses that we are liable to make will be based on this prior probability distribution. If we predict that Tom will form some particular self-interested desire, this will then raise the probability of certain hypotheses about Tom's actions. Trait-attribution thus has a cascading effect on the kinds of mental-states that we attribute, and ultimately on action-prediction.

[35] This is one way in which character traits may serve as an inferential heuristic: without this overhypothesis, mindreaders would begin their action-predictions with a flat probability distribution over all the mental state hypotheses consistent with their current behavioral observations, which would give rise to an inverse problem. Instead, trait attributions bias the prior probability distribution towards a subset of mental-state hypotheses, which the predictor can proceed to test. Even if this distribution is in fact erroneous, it still serves as a means of bootstrapping our initial mental-state predictions, which then allow us to update our priors accordingly.

How we predict that Tom will act on this initial desire-attribution will also be affected by other psychological and situational factors besides Tom's character. For instance, if there are people watching him, we may predict that when Tom looks around, he will refrain from further action. Likewise, if he sees the woman suddenly turn back, we might predict that he will form a new plan: to act as though he were intent on returning the wallet all along, and hand it back to her. Thus, the effects of trait-attributions on specific action-predictions are likely to be moderated by actively updating perspective-taking and belief-attribution procedures that respond to immediate situational factors (Kovács, 2015; Michelon & Zacks, 2006). Importantly, this shows that knowing both 1) that Tom is dishonest, and 2) that someone has dropped a wallet in front of him does not lead to any particular behavioral prediction. Rather, action predictions are produced via an initial trait attribution followed by a series of mental-state inferences at lower levels in the hierarchy.

Conversely, if we antecedently believed that Tom were an honest person, then we might attribute to him the desire to return the wallet to its original owner. Ironically, this might generate a similar series of predicted behaviors as in our original prediction: looking around, reaching down to pick up the wallet, and then giving it back to the woman when she returns, or else pocketing it in order to bring it to the police later. Given two opposite trait attributions in the same situation, there might be no difference whatsoever in the actual behaviors initially predicted; all that would differ would be the kinds of intervening mental states that we ascribe to Tom. Within the pluralists' association-based model, we could never capture this

difference. But on a hierarchical predictive model where traits inform mental-state attributions, which in turn inform predictions about behavior, we can.

One might object that these toy examples only really show that character traits help us to predict desires, but that they don't seem to help us predict beliefs. But there are several ways in which trait attributions might make belief-hypotheses more probable. Traits relating to epistemic agency, such as gullibility or suspiciousness, will affect the priors we assign to hypotheses about beliefs formed on the basis of testimony, for example. Other traits may lead to predictions about how ambiguous situations will be interpreted: we may infer that a paranoid individual will interpret two people whispering one way, an easygoing person another. And even when traits only lead to desire attributions, these might generate expectations about how a person will allocate their attention, which would in turn result in new perceptual beliefs, as when Tom saw the wallet, and then looked around to see if he was being watched.

This hierarchical, predictive approach to trait attribution also helps us make sense of some of the empirical data on trait attribution described in section 2. For instance, if trait attribution is higher up in the action-prediction hierarchy, and has a cascading effect on lower-order mentalistic and behavioral predictions, then we should expect that upon encountering someone new, trait attribution should be prioritized. The more quickly we start to construct a model of a person's character, the faster we will be able to use that information to predict and interpret their behavior. This means that within milliseconds of encountering someone, we need to start to gather whatever information is available to build up a representation of their

stable character traits. Facial structure is well-suited for this purpose, because it can be processed extremely rapidly as coarse-grained, low spatial frequency information (Moshe Bar et al., 2006). This kind of input can be used for an initial conceptual categorization of a stimulus, which can then be used to generate predictions about subsequent input (Moshe Bar, 2007; Chaumon et al., 2014). In other words, we can use facial information to form our first impressions of a person's character, which can then inform our subsequent expectations about intentional behavior.

Of course, initial trait attributions based on faces are neither accurate nor particularly informative for predictive purposes. But in a HPC framework, this is not a problem: learning from mistakes is what Bayesian systems do best. If an initial model results in a prediction error, this information can be used to update the model accordingly. For instance, if one encounters a person with a facial structure associated with trustworthiness, one might initially expect her to be generally well intentioned. However, if one witnessed such a person do something obviously cruel (e.g. abusing an animal), one would of course update one's model of that person's character (Mende-Siedlecki et al., 2013; Tannenbaum, Uhlmann, & Diermeier, 2011). As we accumulate new information about a person's behavior, we may iteratively revise our initial model of their character, leading to increased accuracy (Cunningham, Zelazo, Packer, & Van Bavel, 2007). This helps to resolve one of the major puzzles of the pluralist account: even if trait attributions are initially unreliable, they might still serve as a basis for social learning and prediction, leading to increasingly accurate models of a person's character.

This complementary prioritization and updating of trait information can also shed light on the cognitive processing underlying the fundamental attribution error. Most impression-formation tasks that lead to the fundamental attribution error introduce participants to new people. If constructing a character model is prioritized by the mindreading system, then we should expect interpreters to use whatever behavioral information is available to construct that model as fast as possible. But when our attention is drawn to mitigating situational factors, our initial personality models are updated, and the behavioral evidence is discounted (Gilbert et al., 1988).[36] Likewise, when we are provided with additional mental state information (Krull et al., 2008; Reeder et al., 2004), or primed to think about mental states (N. Hooper et al., 2015), we use that information to update our character models accordingly.

However, in an HPC framework, not all prediction errors lead to updating. The world, after all, is messy and complex. Even a highly accurate causal model is liable to make mistakes. Many of these mistakes will be due to noise in the input, rather than a problem with the model. Updating the model to accommodate every piece of information it encounters would result in overfitting, and thus diminish its overall predictive accuracy (Hohwy, 2013). Moreover, updating the model is likely to be cognitively costly, since it would require additional memory searches and generative procedures. Updating, in other words, can be a bad thing. Sometimes, prediction errors ought to be discounted as noise.

[36] Members of "collectivist" cultures, who habitually attend to contextual factors, no doubt benefit from such attentional effects in their comparative resistance to the correspondence bias.

Modeling character traits is no different. One might have a fairly accurate representation of a person's character, and still occasionally be surprised by their behavior. For instance, one might be quite surprised to learn that Adolf Hitler was a vegetarian. Such information could be used to update one's model of his moral character, but this seems unlikely. Rather, one would simply ignore this information, and continue to rely on one's prior model. But this raises an interesting question: when do we update our character models in the face of new, conflicting behavioral information, and when do we treat it as noise?

A recent study by Daniel Ames and Susan Fiske hints at an answer (D. L. Ames & Fiske, 2013). Given that updating character models is likely to require the use of additional, limited cognitive resources, they hypothesized that people should selectively allocate those resources towards targets that are most behaviorally relevant to them. To test this hypothesis, they first introduced subjects to two confederates, who were described as "expert consultants" with whom the participants would later collaborate with in a joint project after first performing a solo task. Participants in the outcome-dependent condition were told that based on their performance in this joint task, they would be considered for a $50 prize. Participants in the outcome-independent condition, in contrast, were told that their eligibility for the prize would be based on their performance in the solo task only. Subjects then underwent fMRI scanning and were shown statements about the two confederates that were either consistent or inconsistent with what they had previously been told about them.

Ames and Fiske found inverse patterns of activity in the dmPFC for the two conditions: participants in the outcome-dependent condition displayed more responsivity to inconsistent information, whereas participants in the outcome-independent condition showed more responsivity to the consistent information. The authors suggest that the outcome-dependency manipulation led participants to use different updating strategies: when achieving their goal (the reward) depended upon interacting with the confederate, they used the inconsistent information to update their character model, so as to better predict and adjust to their partner's behavior. In the outcome-independent condition, in contrast, participants tended to dismiss the inconsistent information as noise, and thus conserve cognitive resources.

From a HPC perspective, we can interpret this outcome-dependent updating as reflecting higher-order predictions about the action-relevance of a prediction error. When surprising information is particularly important for action planning (e.g. when we expect to interact with a person in the future), we are more likely to incorporate that information into our predictive models, instead of dismissing it as noise. On the other hand, when a bit of surprising information is not action-relevant (e.g. when we do not expect to interact with a person in the future), we are less likely to devote resources to updating our predictive models, and more likely to dismiss the prediction error as noise. In other words, when a prediction error is more relevant to our goals, we "raise the volume" on that signal; when it is less relevant, we "turn the volume down." In neuro-cognitive terms, this modulation of expected precision translates into shifts in attention. This may explain why participants under cognitive load are more likely to fall prey to the fundamental attribution error: when their attention is directed

towards another task, they fail to attend to update their character models in response to error signals coming from situational information.

In sum: temporally stable character traits are represented at the upper level of an action-prediction hierarchy, and are used to generate prior probability distributions for hypotheses about more transient mental states, including beliefs and desires. These hypotheses are then used to inform hypotheses about even more transient states, which are in turn used to predict or interpret behavior. The downstream effects of trait-attributions on action-prediction are liable to be modulated by active belief-attribution procedures operating at lower-levels in the hierarchy. Prediction-error signals are conveyed back up the hierarchy, and are either used to revise the model at the appropriate level, or dismissed as noise. The action-relevance of an input can modulate whether prediction errors are treated as noise or used to revise the model by changing the expected precision in of an input signal.

6. Future directions

Adopting an integrative hierarchical approach to reasoning about character traits enables us to make a number of novel predictions. Broadly speaking, we should expect that manipulating background information about a person's character should lead to differences in the kinds of mental states that we attribute to them, especially when interpreting ambiguous actions. More specifically, manipulating trait attributions along the warmth dimension should lead to either more negative or more positive desire and intention attributions. For instance, individuals presented as low-warmth should be interpreted as having harmful or self-serving desires, while

individuals presented as high warmth should be interpreted as having helpful or altruistic ones (as we saw in section 2, this kind of effect is likely to be responsible for the side-effect effect (Sripada & Konrath, 2011; Sripada, 2012)). Manipulations along the competence dimension should lead to differences in the amount of knowledge that we attribute to them: more competent individuals should be more likely to be viewed as having the appropriate beliefs and making the appropriate inferences, while less competent individuals should be more likely to be viewed as ignorant. Warmth and competence information could be conveyed through facial information, through the observation of diagnostic behaviors, interactive experiences, or through testimony.

Beyond these initial predictions, this account of character-trait attribution offers us a new way to connect the study of mindreading with our understanding of stereotyping, prejudice and implicit bias (Spaulding, 2016). As was noted earlier, the contents of common stereotypes are characterized by variation along the same two dimensions as ordinary trait attributions, suggesting that we use cues of group membership to infer that an individual will possess a particular set of character traits (e.g. an elderly person will be viewed as kind, but also as incompetent). Cues to group membership, such as skin color or accent, thus seem to play a similar role as facial features in conveying trait information; however, stereotypes seem to contain clusters of trait information, rather than just single traits. If trait attribution affects mental-state attribution as I've described, and stereotypes allow us to attribute clusters of character traits to individuals, then it would follow that stereotypes should also affect mental-state attribution.

As it happens, there is already some evidence that this is the case. Sagar and Schofield showed sixth-grade children vignettes depicting ambiguously aggressive dyadic interactions between students, such as one student bumping into another in a hallway, asking for food in the cafeteria, poking another student, and taking a pencil without asking. The authors also systematically manipulated the race of the actor in each dyad, such that some participants saw a white actors bumping, asking for food, poking, etc., while others saw black actors doing so. They found that the behaviors of black actors were interpreted as more mean and threatening than the identical behaviors from white actors (Sagar & Schofield, 1980). Thus, when the intention underlying an action is ambiguous (e.g. intentionally threatening and aggressive versus neutral), observers fell back on stereotypes about black aggressiveness to interpret it. These results suggest that ascertaining the role of character-trait attribution in mindreading may also help us to better understand the cognitive basis for implicit racism.

The connection with stereotyping also raises the possibility that, in addition to their role in action-prediction, trait attributions may also serve a social function.[37] One of the explanations that has been offered for why we represent traits along the warmth/competence dimensions is that warmth helps us to keep track of potential threats, while competence helps us to keep track of agents' social status (Fiske et al., 2007). Notably, threat and status are not intrinsic properties of individuals, but rather relational, social properties that tend to vary with context: who counts as threatening or high status often depends upon one's own group identity and social rank. Thus,

[37] Thanks to an anonymous reviewer for suggesting this.

while we tend to represent traits as intrinsic, stable properties of individuals, it may be that what we are really tracking are social relationships.[38] These factors will still be highly relevant to action-prediction, however: whether an individual is higher or lower-status than us, or a member of the in-group or out-group, will have a significant effect on how they decide to act towards us. This may explain the predictive utility of trait-attributions: even if the situationists are right, and stable character traits do not really exist, we can still use trait representations as a proxy to help us factor social identity into our action-predictions.

7. Conclusion

Integrating character-trait attribution into a hierarchical Bayesian account of theory of mind promises to enrich our understanding, not just of these two sets of phenomena, but also a network of related phenomena of substantial social and philosophical importance. Traditional accounts of mindreading have paid little heed to character-trait attribution, focusing instead on the attribution of beliefs and desires. Folk psychology pluralists have rightly pointed this oversight, and taken important steps towards drawing attention to the significance of trait reasoning in folk-psychological inference. However, the pluralist account treated trait reasoning as a completely independent form of behavioral prediction, which does not fit well with the empirical data. In contrast, I have argued that the attribution of character traits is systematically related to the attribution of other forms of mentality, and that a hierarchical Bayesian

[38] This may be an instance of what Cimpian and Salomon call the 'inherence heuristic' – a "fast, intuitive heuristic leads people to explain many observed patterns in terms of the inherent features of the things that instantiate these patterns" (Cimpian & Salomon, 2014, p. 461) – and a precursor to psychological essentialism about certain social categories (Gelman, 2004; Haslam, Bastian, & Kashima, 2006; Rhodes, Leslie, & Tworek, 2012).

architecture is a promising way to explain that relation. This account yields a number of novel empirical predictions about mindreading, and also has the potential to further unify the study of mindreading with neighboring empirical domains, such as stereotyping and implicit bias.

Appendix: Pragmatic development explains the theory-of-mind scale[39]

1. The nativist–constructivist debate

Humans are hyper-social. This much is widely agreed. It is also generally agreed that human social *cognition*—involving a capacity to attribute mental states to other people and to anticipate their likely actions—is essential to human uniqueness (Tomasello, 2009), even if it isn't the ultimate source of that uniqueness (Piantadosi & Kidd, 2016). Accordingly, a great deal of effort has been expended over more than 30 years in an attempt to understand the development of human mindreading capacities (Wimmer & Perner, 1983). For most of this period there was a widespread consensus that such capacities are constructed gradually over the course of the preschool years, relying on linguistic and cultural input together with general-learning and theorizing abilities (Gopnik & Wellman, 1992; Wellman et al., 2001; Wellman & Woolley, 1990). While there were always some in the field who claimed that basic mindreading abilities are innate, and that the appearance of development reflects failures of performance (Scholl & Leslie, 1999), this was decidedly a minority position.

In the past 10 years, however, the field has changed dramatically. There are now dozens of studies of infants aged 6 to 18 months using a variety of non-verbal methods (including expectancy-violation looking, anticipatory looking, active

[39] This Appendix was written collaboratively with Peter Carruthers. It was originally published as Westra and Carruthers (2017), and has been reprinted here with permission from Elsevier, copyright license # 4067641072972.

helping, and more) suggesting that infants understand the goals and beliefs of other agents, and can anticipate actions accordingly. (For example, see: D. Buttelmann et al., 2009b, 2014; F. Buttelmann et al., 2015; He, Bolz, & Baillargeon, 2012; Kovács et al., 2010; Onishi & Baillargeon, 2005b; Southgate & Vernetti, 2014.) It is now widely agreed that these findings reflect an underlying social-cognitive competence of some sort (although see Heyes (2014a) for a dissenting view). What is disputed is how these early abilities relate to those that underlie performance in more traditional verbal tasks. Nativists have seized on the new findings to claim that core mindreading abilities are present throughout infancy, and that early failures on verbal tasks reflect performance difficulties of some sort (Baillargeon et al., 2010; Carruthers, 2013). Constructivists, in contrast, have mostly converged on some form of two-systems view, according to which there is an early-developing, implicit, limited-flexibility system that is later supplemented by a slowly-acquired, flexible and explicit, theory of mind (Apperly, 2011; Wellman, 2014).

There are broadly two lines of support for this new constructivist position. One consists of evidence that both implicit and explicit systems exist alongside one another in adults, and that the implicit system operative in infancy has signature limits (Apperly & Butterfill, 2009; Apperly, 2011; Butterfill & Apperly, 2013; Schneider, Bayliss, et al., 2012; Schneider, Slaughter, & Dux, 2014). This evidence has been systematically criticized elsewhere (Carruthers, 2015a, 2015c; Christensen & Michael, 2015; Thompson, 2014; Westra, 2016b). The other line of support derives from evidence of an orderly and systematic progression in toddlers' verbally-manifested mindreading abilities, which is suggestive of genuine conceptual

development. This is most clearly demonstrated by Wellman and colleagues who have created and validated across cultures the *mindreading scale*. This will be our main focus here. Our goal is to show that the data provided by the mindreading scale fail to support constructivism. This is because there are plausible alternative explanations —mostly pragmatic in nature—that have not yet been controlled for and excluded.

2. The mindreading scale

Wellman and Liu undertook two studies (Wellman & Liu, 2004). The first was a meta-analysis of investigations of mindreading development in which children's understandings of different types of mental state were pitted against one another using otherwise-matched tasks. (All of the studies reviewed involved verbal presentations and required the children to give verbal answers.) Their analysis showed that the first milestone children pass is understanding that different people can have different desires, and that these differences will lead them to act differently. These tasks are reliably easier than ones in which children are required to understand that different people can have different beliefs. The latter tasks are in turn easier than ones in which children are required to understand that someone can be ignorant of a fact by virtue of lacking perceptual access to it. Finally, understanding ignorance is reliably easier than understanding that people can have, and act on, beliefs that are false.

Inspired by these meta-analytic findings, Wellman & Liu (2004) constructed a sequence of matched tasks, extended to include a test of children's ability to

understand that someone can act in a way incongruent with her true feelings.[40] They included a diverse-desires task (DD), a diverse-beliefs task (DB), a knowledge / perceptual-access task (KA), a false-belief task (FB), and a hidden-emotions task (HE). They tested 75 children aged 3–5 on all of these tasks, finding evidence of a robust developmental progression that matched the meta-analytic findings, with an understanding that people can hide their true emotions being hardest of all. In fact, a large majority of the children performed in a manner consistent with the following order of ease of passing: DD > DB > KA > FB > HE. Since Wellman & Liu's initial study, over 80% of some 500 children tested in the USA, Canada, Australia, and Germany have displayed abilities consistent with this pattern (Kristen, Thoermer, Hofer, Aschersleben, & Sodian, 2006; Wellman, 2012). Moreover, congenitally deaf children born of hearing parents (who are introduced to full-blown sign-languages much later in childhood than normal) follow the same developmental progression, only significantly delayed (Peterson et al., 2005; Peterson & Wellman, 2009).

Wellman and colleagues have also found that this developmental sequence is cross-culturally robust, with one intriguing exception: preschool children from "collectivist" cultures (specifically, China and Iran) tend to find the knowledge-access (KA) task easier than the diverse beliefs (DB) one, thus exhibiting the sequence DD > KA > DB > FB > HE (Duh et al., 2016; Shahaeian, Peterson, Slaughter, & Wellman, 2011; Wellman, Fang, Liu, Zhu, & Liu, 2006). This is thought to reflect a cultural emphasis on differences of opinion in "individualist" countries such as the USA, and a correspondingly greater emphasis on education, knowledge,

[40] Some of the tasks included in Wellman & Liu's (2004) initial battery of tests were dropped from follow-up studies, and will not be discussed here.

and the importance of learning from those in authority in "collectivist" ones.

In addition, Rhodes & Wellman (2013) combined use of the mindreading-scale tasks with microgenetic measures (a form of longitudinal study in which behavior is sampled very frequently, which effectively amounts to a form of training). Children in the study were pre-tested on the mindreading scale, and those in the experimental condition then underwent a number of regular microgenetic training sessions over the course of six weeks. In each of these sessions children had to complete two new false-belief prediction tasks. They were then shown the correct outcome of the scenario, and were asked to explain the character's action. Consistent with previous intervention studies (Amsterlaw & Wellman, 2006; Lohmann & Tomasello, 2003), training on false-belief tasks tended to have a positive effect on performance at post-test. More interestingly, it was also found that children's scores on the mindreading scale at pre-test predicted the effectiveness of the training. Children who could already pass the knowledge-access task at pre-test were more likely to pass the false-belief task at post-test than children who could only pass the diverse-beliefs task at pre-test. Using similar methods, Wellman and Peterson obtained comparable training effects for older late-signing deaf children (Wellman & Peterson, 2013).

Wellman (2012, 2014) argues that this overall body of data supports a constructivist account of mindreading development, and is correspondingly problematic for nativist theories. Children are said to be constructing a causal framework for understanding the operations of the mind, drawing on their own experiences and their observations of others. Some aspects of the developing theory

(particularly the idea that the mind contains states that *represent* aspects of reality, needed for an understanding of false belief) are said to be intrinsically harder to construct than others. But construction of the theory also depends on cultural input. Those who are on the cusp of constructing a full-blown representational theory of mind are most likely to transform intensive conceptually-relevant forms of social experience into full false-belief competence, but such experience still benefits children at an earlier stage in the mindreading-scale progression. In contrast, if mindreading capacities are innate, then it is said to be very unclear why performance should exhibit these regularities, or why cultural differences and individual training-experiences should make any difference.

Wellman draws a false contrast here, however. For nativism is consistent with cultural learning. What is innate, it can be said, is a domain-specific learning mechanism. (Compare what nativists say about the innateness of the language-faculty, which is obviously designed for learning.) Specifically, a nativist can claim that infants are innately endowed with certain core concepts (perhaps DESIRE, BELIEF, PRETENSE, HAPPY, SAD, SEE, and TELL) and certain basic principles of attribution (such as "seeing leads to believing"). Thereafter novel concepts can be acquired, and new principles of attribution learned, relying both on individual experiences and cultural input. So from this perspective it isn't surprising that culture might make a difference, nor that training might help performance. Moreover, it may be that the kind of learning that actually contributes to passing the tests making up the mindreading scale doesn't require enrichment of the target mental-state concepts at all. Rather, as we will see, it may be a matter of learning to recognize cues that signal

the current topic of conversation or the most likely intent behind a question.

In addition, it is far from obvious that Wellman's own constructivist framework is internally coherent. Specifically, it is unclear that the delay between an ability to pass diverse-belief tasks and a capacity to pass false-belief tasks makes theoretical sense, from a constructivist perspective. This is because both tasks require a grasp of the representational nature of mind. In order to understand that two people can have different beliefs about the same subject matter, one needs to understand that the subject matter in question can be represented differently. But this is the *same* understanding as has often been thought to underlie grasp of the possibility of false belief, together with the ability to pass (verbal) false-belief tests. Moreover, since the two beliefs in a diverse-belief scenario conflict with one another, at least one of them must be false.

Of course it is true that in some versions of diverse-belief test the child only guesses at the location of the item, rather than seeing it for herself. But at the very least we can say that the diverse-belief test requires the child to reason about what someone will do who has a belief that conflicts with what the child has just *said* she thinks is the case. Why should this be any easier *conceptually* than reasoning about what someone will do who has a belief that conflicts with what the child has just been *told* is the case (which is what happens in the mindreading-scale version of the false-belief test)? Of course, if guesses give rise to beliefs of lesser strength than testimony from an adult, it may be that the pre-potent response (the "lure of the real") in a false-belief task is correspondingly stronger. But this would then suggest that the differential in performance reflects differences in executive function, rather than

differing understandings of the mind. So there is still a problem, here, for Wellman's preferred conceptual-development interpretation of the mindreading scale.

From the fact that Wellman's constructivist framework is problematic it doesn't follow that a nativist account is correct, of course. Indeed, the problem for nativists arising from the mindreading-scale data is that the very experiments with infants that are thought to support nativism suggest that infants already possess the concepts and attribution-principles tested by most of the items on the scale (Baillargeon et al., 2010; D. Buttelmann et al., 2009, 2014; Luo & Baillargeon, 2005). Moreover, although the infancy-data seemed initially to suggest that the development of infant competence might mirror two major stages of the mindreading scale (specifically, that desires are understood earlier than beliefs), recent data puts infant capacities to track and reason about false beliefs as early as 7 or 8 months, or even 6 months of age (Kampis, Parise, Csibra, & Kovács, 2015; Kovács et al., 2010; Southgate & Vernetti, 2014). So there is little scope, here, for arguing that mindreading-scale performance merely lags behind true mindreading development, requiring monotonic growth in executive-function capacities or linguistic abilities, for example. The challenge for nativism, then, is to explain why the mindreading scale should be so robust if all the conceptual resources necessary to succeed in the tests are available some two years earlier than children actually begin passing its easiest items.

Not all tasks in the mindreading scale contribute to this challenge for nativism, however. Specifically, the final "hidden emotion" (HE) item is rather different in character from the rest. This is so for two reasons. The first is that the

story presented to children is more complex than gets used with the other measures in the scale, and more memory-check questions get asked prior to the target question. So the task is likely to be significantly harder for reasons extraneous to mindreading. Second, the HE task tests an appreciation of how mental states, on the one hand, and behavior that would normally manifest such states, on the other, can be in conflict, whereas the belief and desire tasks are about how the states of different *people* can conflict. It may be that it takes a while to acquire the knowledge that people aren't always feeling what they appear to be feeling. And crucially for our purposes, there is no evidence that infants have any sort of appreciation of this point. So there is no initial puzzle here for nativists to answer.

Accordingly, in what follows we will consider just the first four items of the mindreading scale (diverse-desire, diverse-belief, knowledge-access, and false-belief). We will first focus on explaining the sequence DD > DB > FB, before turning to a separate discussion of KA. Our goals are (a) to provide well-motivated alternative explanations of the reliability of the sequence DD > DB > FB, together with (b) the influence of culture on the order in which children pass DB and KA, as well as (c) the boost that false-belief training can give children who perform at intermediate levels in the sequence. Our goal is not to demonstrate that our alternative explanations are correct, however. That would require a whole raft of new experiments. Rather, it is to show that they are independently plausible, thereby undercutting any support for constructivism from the mindreading-scale data in the absence of such experiments.

3. Existing accounts of verbal-task performance failures

Since nativists are committed to claiming that the conceptual–theoretical competence for passing all the main components of the mindreading scale are present from infancy onwards, they must explain the mindreading-scale findings in terms of differential demands on performance. What resources are available for constructing such an explanation? We will first consider what nativists might say about the failures of children younger than four to pass verbal false-belief tasks, despite passing non-verbal versions of the same tasks from as early as the latter half of the first year of life.

One possibility often mentioned in the literature concerns executive-function abilities. It has been said, for example, that capacities to pass verbal false-belief tasks depend on late-maturing fronto-parietal pathways, connecting executive function in the frontal lobes with major components of the mindreading network (Baillargeon et al., 2010). In support of such a view, executive function is known to correlate with age of passing verbal false-belief tasks (Carlson et al., 2002; Carlson & Moses, 2001; Kloo & Perner, 2003). Moreover, reduced demands on executive function are thought to explain why removal of the target object from the scene in a change-of-location false-belief task makes the task somewhat easier (Wellman et al., 2001; Southgate et al., 2010).

While executive function abilities are no doubt part of the story, they can by no means provide the whole explanation. There are a number of reasons for this. One is that the correlation between executive function and false belief is small after controlling for age and verbal ability (only .22; see Devine & Hughes, 2014). Another

is that the active-helping false-belief tasks passed by 18-month-old infants surely require executive decision making (Buttelmann et al., 2009, 2014, 2015). Moreover, although Chinese children are known to be more advanced than US children in their executive-function abilities, they perform no better in false-belief tasks (Sabbagh, Xu, Carlson, Moses, & Lee, 2006; Wellman et al., 2001). And in addition, it is hard to see how an appeal to executive function can explain the increasing difficulty of the tasks that make up the mindreading scale, which are generally well-matched in terms of their executive demands. Nevertheless, executive function will surely assist with learning and managing the pragmatic aspect of verbal mindreading tasks, which we emphasize below.

Another factor known to correlate with age of passing false-belief tasks is general language ability. A number of constructivist accounts have proposed that the acquisition of language plays an important, perhaps necessary, role in the development of mindreading. However, there is substantial disagreement about which aspects of language play this role. Some authors have suggested that it is complementation syntax (de Villiers & Pyers, 2002), others have emphasized mental state vocabulary (Montgomery, 2005), and yet others stress the social experience that comes from linguistic interactions (Tomasello & Rakoczy, 2003; Dunn & Brophy, 2005; Harris et al., 2005). But in their meta-analysis Milligan et al. (2007) were unable to identify a special role for any single aspect of language independent of general language ability. And in all, after controlling for age, they determined that linguistic factors correlated only moderately with mindreading ability (.31), and accounted for only 10% of the variance in the latter. Moreover, general language

ability is often controlled for in testing the validity of the mindreading scale (Wellman & Peterson, 2013). It is therefore unclear how a simple appeal to language ability could explain the sequential progressions in mindreading performance described earlier. What is surely correct, however, is that what makes verbal false-belief tasks hard has *something* to do with the fact that they are verbal (or at least communicative) in nature. The proposal we make below will build on this idea.

Yet another suggestion that has been made in the literature is that verbal false-belief tasks are, in effect, *triple*-mindreading tasks (Carruthers, 2013). This is because both language comprehension and communicative production are inevitably pragmatic in nature, and because it is widely acknowledged that mindreading is implicated in pragmatic aspects of speech. The child in a verbal false-belief task has to figure out what the experimenter is trying to communicate while computing, remembering, and accessing at the relevant time the mental states of the target agent. The child then has to figure out and select a verbal (or other communicative) response to have the intended effect on the mind of the questioner. A verbal false-belief task will thus involve a complex interplay between executive decision making, the language faculty, and mindreading. It seems plausible that the relevant connections (and the efficiency of the mindreading system itself) might not have matured sufficiently in younger children for them to pass. Although promising, however, this suggestion is not specific enough to explain the comparative difficulty of the tasks that make up the mindreading scale.

Pragmatic accounts of young children's failures in verbal false-belief tasks are not new. Siegal & Beattie (1991), for example, hypothesized that children might be

misinterpreting the test question to mean, "Where *should* she look for her ball?" They found that by altering the question slightly to, "Where will she look *first* for her ball" they were able to shift the average age of passing a few months earlier. (The latter question doesn't remove the ambiguity altogether, of course. It can still be heard as asking, "Where *should* she look first for her ball [in order to get it right away]?") Surian & Leslie (1999) later replicated this finding, while also showing that the "look first" manipulation has no effect on children with autism (suggesting that the difficulties these children experience with false-belief tasks are not merely pragmatic). More recently, Helming et al. have proposed a more elaborate form of pragmatic account, which we briefly discuss here before developing our own view in Section 4 (Helming et al., 2014; Helming, Strickland, & Jacob, 2016).

Helming et al. argue that young children have problems with the false-belief task because it requires them to adopt two different perspectives simultaneously. In order to pass, a child must adopt a third-person—"spectatorial"—perspective on the protagonist's instrumental action, while simultaneously adopting a second-person—communicative, and hence cooperative—perspective with the experimenter and when answering the latter's question. Engaging with the experimenter in a communicative interaction is said to disrupt the child's third-personal tracking of the protagonist's beliefs. The child's subsequent response is then the product of two pragmatic biases: one referential, one cooperative.

The referential bias is triggered when the experimenter asks a question about the target object. It can be triggered by the test question itself (e.g. "Where will she look for her ball?"), or by a prior control question about the actual location of the

target object (e.g. "Where is the ball now?") (Rubio-Fernández & Geurts, 2015). Such questions have two primary effects: first, mentioning the object primes the child to think about its true location. Second, involvement in a second-person interaction with the experimenter causes the child to focus on their shared epistemic perspective (specifically, their shared knowledge of the object's true location). This disrupts the child's ability to track the protagonist's false belief from a third-person perspective. Together, these two factors cause the child to focus on the true location of the object, and ignore the agent's false belief.

The cooperative bias is said to arise from the fact that children are motivated to help the mistaken agent. This leads them to adopt a second-person perspective toward the protagonist in the narrative, rather than a third-person spectatorial one. The bias is a manifestation of the fact that children at this age are chronically helpful, and will go out of their way to help an unknown adult even when it takes effort to do so and they are engrossed in an activity of their own (Warneken & Tomasello, 2007, 2009, 2013; Warneken, 2015). Indeed, even somewhat younger children will point out information to help an ignorant adult who is searching for something, or to prevent an adult from making a mistake (B. Knudsen & Liszkowski, 2012; Liszkowski, Carpenter, & Tomasello, 2008). Thus, when younger children see that the protagonist in the false-belief task is mistaken about the location of her ball, they are motivated to help her find it. This, in combination with the referential bias towards the true location of the ball, leads the child to misinterpret the experimenter's predictive question, "Where will she look for her ball?" as a normative one ("Where *should* she look for her ball?")

We agree with much in these suggestions, so far as they go.[41] But while these two biases may help to explain why younger children initially *fail* the false-belief task, they do not explain how older children eventually come to *pass* it. Here, Helming and colleagues appeal to children's developing executive abilities (which presumably help children inhibit the two biases).[42] But for the reasons that we have just mentioned, executive development cannot be the whole story. Moreover, Helming and colleagues' account is silent about the ordered difficulty of the tasks that make up the mindreading scale. In Section 4 we will construct a more elaborate pragmatic account of false-belief failures that incorporates this one, building on the work of Westra (2016a).

4. A pragmatic account of false-belief performance

We should stress at the outset that all communication is inevitably partly pragmatic in nature. A communicator produces a performance of some sort (a speech act, a gesture) and the audience has to figure out the intent behind that performance (Sperber & Wilson, 1995, 2002). Moreover, recent models of speech comprehension suggest that it takes place competitively and in parallel, with syntax, semantics, and pragmatics being processed interactively, generally with multiple hypotheses in play

[41] In their account, Helming et al. (2014, 2016) stress that these two biases are generated by the demands of simultaneously adopting second- and third-person perspectives on the actions of the experimenter and the agent. But we are doubtful whether this framework is really doing any work. The referential bias seems to be generated primarily by the fact that the hidden object has been mentioned, while the cooperative bias is the product of children's disposition to engage in helping behavior. It isn't obvious what pointing out the contrast between second- and third-person perspectives adds to explanation. Therefore, although we agree with the substance of Helming et al.'s account, we do not follow them in adopting this terminology.

[42] While they largely endorse an executive-functioning account in their paper, Helming and colleagues do note that the existence of the two biases does not logically entail acceptance of an executive-functioning account. This, they acknowledge, might make their pragmatic analysis of the task compatible with some form of constructivism. By the same token, it also makes their view compatible with our developmental proposal.

at each level (Hickok & Poeppel, 2007). In production, too, it seems that speakers make a selection from among a number of candidate utterances suggested by the context (Novick, Trueswell, & Thompson-Schill, 2010). We should expect, then, that a child participating in a verbal false-belief task will be no different. A number of candidate interpretations of the experimenter's question are likely to be entertained and evaluated for likelihood (albeit swiftly and unconsciously), with a selection from among candidate answers being made accordingly.

How a child interprets the experimenter's questions in a false-belief experiment will depend, in part, on her construal of the nature of the communicative exchange: that is to say, its topic and purpose. As Helming et al. (2014, 2016) rightly point out, one aspect of the false-belief scenario that will seem highly salient to children of this age is the fact that the agent is in need of help (because she has a false belief). If nativism is assumed, then the child will be aware that the protagonist *needs* help, and may thus anticipate being invited to offer such help. This will, in turn, increase the salience of a normative interpretation of the test question, taking it to mean, "Where *should* she look for her ball?" Even more simply, the child might infer that the experimenter is inviting her to help Sally find her ball, and interpret the test question as, "Can you show Sally where to look for her ball?" We will refer to this as the "helpfulness-interpretation" of a false-belief (or other mindreading-related) question.

Another salient construal of the communicative exchange with the experimenter is that it serves a pedagogic purpose of some sort. Hence the false-belief question is a request for the child to show what she has learned from the exchange;

she is being asked to exhibit that she has acquired some target item of knowledge (whatever that is). We will refer to this as the "knowledge-exhibiting-interpretation" of the question. In fact, children are quite likely to assume that the interaction with the experimenter may have a pedagogic intent. This is because the normal cues to pedagogy (shared attention, eye-contact between adult and child at the outset of the exchange) will almost always be present in a normal false-belief experiment. Note that in other contexts such cues have been found to serve as reliable signals to young children that knowledge transmission of some sort is about to take place (Csibra & Gergely, 2009; Gergely, 2013).

Now notice that intended interpretation of the false-belief question is, indeed, a knowledge-exhibiting one: the experimenter wants the child to exhibit her knowledge of the psychological states of the protagonist in the story and/or their likely effects on behavior. And this will be the interpretation most directly suggested by the syntax and literal semantics of the question ("Where does she think it is?" / "Where will she go?") But this interpretation requires the child to take the topic of conversation to involve the protagonist's cognitive states of knowledge or belief. This may strike the child as unlikely, for reasons we will explain shortly. Moreover, there is normally nothing in the setup of a false-belief experiment to suggest to the child that she is supposed to be learning something about the cognitive states of the protagonist (although manipulations that make cognitive states more salient do have the effect of reducing the age at which children first pass the false-belief task; see Wellman et al., 2001).

In contrast, what will seem most immediately salient about the false-belief

scenario is that it involves displacement and concealment of an object (in a change-of-location false-belief task), or that a container has contents other than one might expect (in an unexpected-contents version of the task). If pedagogic intent is assumed, then these will seem like probable targets for learning: the experimenter wants the child to learn about the true location of the object or the contents of the container. The child is then likely at least to entertain the hypothesis that the question is inviting her to exhibit her knowledge of the worldly events that have just unfolded—that is, the actual location of the object, or the actual contents of the container. As a result, in false-belief experiments there will generally be two knowledge-exhibiting interpretations in play, in addition to the helpfulness-interpretation discussed earlier. The child's task is to figure out which of the three is the most likely.[43]

When a child in a false-belief experiment is asked where the protagonist will look for her ball, then, we suggest that there will generally be three interpretations of the question that are activated, competing to control the answer. One is that the child is being invited to be helpful toward the protagonist. Another is that she is being asked to exhibit her knowledge of the events that have unfolded in the story. And the third is that she is supposed to exhibit her knowledge of the way in which the protagonist's beliefs will issue in action. Notice that although this third interpretation is the one intended by the experimenter, each of the others will push in the direction of the same (incorrect) answer: both will incline the child to reply by stating the

[43] Consistent with this suggestion, Howard et al. (2008) found that the most consistent predictor of false-belief performance in their corpus data was the frequency of child-directed questions. Likewise, Hughes et al. (2014) found in a study employing false-belief tasks with slightly older children from England, Italy, and Japan that the only systematic predictor of differential success across groups was the age at which children in their respective countries begin formal schooling. (This is a year earlier in England than in Italy and Japan, and the English children performed significantly better.) For of course school-teachers frequently ask children knowledge-exhibiting questions.

actual, current, location of the ball. One might expect, then, that in a three-way competition among possible interpretations, the odd one out would face an uphill battle to control behavior. Put differently: the child has *two* reasons to name the actual location of the ball, and only one reason to name the location believed-in by the story protagonist.

This account enables us to offer a deeper explanation of the referential bias postulated by Helming et al. (2014, 2016), which was discussed in Section 3. The reason why control-questions or false-belief questions that mention the actual location of the target object are more likely to lead to erroneous answers is that they raise the probability of the two competing interpretations of the intent behind the question. When the experimenter refers to the target object she thereby draws attention to its actual location. This will make that location seem relevant to the communicative exchange, hence increasing the likelihood that she is inviting the child to be helpful by pointing out that location to the protagonist; and by the same token, it will also make it seem more likely that the child is being invited to exhibit her knowledge of the actual location, rather than her knowledge of the protagonist's beliefs. We diverge from Helming et al., however, in that we do not view the referential bias as *disrupting* or *interfering with* the third-person mindreading process. On our account, the child continuously represents the agent's false belief throughout the task. However, children don't use this information when interpreting the experimenter's question, because they are drawn instead to more salient, alternative interpretations.

We noted earlier that there is a systematic reason why it may strike the child as unlikely that the topic of conversation in a false-belief task is the protagonist's

mental states and resulting actions, or that the protagonist's beliefs are conversationally relevant. This is that, in the child's experience, cognitive states are rarely talked about (Westra, 2016a). One reason for this is that our ordinary explanations and descriptions of behavior generally leave beliefs implicit. Instead, we simply refer to an agent's desires, leaving it to our interlocutors to infer the relevant belief-factors (Papafragou et al., 2007; see also Steglich-Peterson & Michael, 2015). Indeed, Papafragou et al. (2007) found that participants would only spontaneously mention beliefs when describing behavior in cases of deception or false belief, or when provided with particular syntactic cues. While mature speakers will recognize that these exchanges contain implicit references to beliefs, a novice speaker unfamiliar with the pragmatics of belief discourse will likely come to regard references to beliefs as relatively rare events. Moreover, such a pattern of omission is reflected in child-directed speech, where "think" gets used only half as frequently as "want" (Taumoepeau & Ruffman, 2006; MacWhinney, 2014).

In addition, we often use verbs like "think" in a manner that isn't really about beliefs at all (Simons, 2007; Lewis et al. 2012). Rather, "think" is frequently used as a way of indirectly asserting its complement. Thus if one says, "I think it will rain this afternoon," one's primary speech act is to make a hedged assertion about the weather, not to attribute a belief about the weather to oneself. Third-person uses of these terms often perform a similar role. If one responds to someone's query about the upcoming weather by saying, "John thinks it will rain this afternoon", mention of John's beliefs is introduced in an evidential role, and the result is still something resembling an indirect assertion that it will rain. The topic is still the weather, not John's mental

states.

Studies of corpus-data collected from children's conversations with adults show that such indirect-assertion uses of "think" make up a large proportion of children's conversational experience with such terms, both in child-directed adult speech and in children's own speech production (Shatz et al., 1983; Bloom et al., 1989; Diessel & Tomasello, 2001). Sentences of the form, "S thinks that P" are more likely to serve as a way of indirectly asserting, "P" than to attribute a belief to the subject. This has led some linguists to argue that children interpret "think" as indirect by default, and only draw on the attributive sense when the indirect interpretation is clearly implausible (Lewis et al. 2012; Hacquard 2014; Dudley et al., 2015).

Thus, while many of our actual thoughts about the beliefs others are left implicit, many of our explicit uses of belief-verbs tend not to be about beliefs at all. All this will lead a novice speaker to assume that conversations about beliefs are quite infrequent. So even when an utterance might be plausibly interpreted as being about beliefs (e.g. "What will she think is in the box?"), these interpretations will be assigned a low prior probability. Consequently, such interpretations are unlikely to be selected when more probable alternatives exist.

It might be objected that not all forms of false-belief task use the term "think", either in describing the false-belief scenario or in the test question. Rather, the child might merely be told where the protagonist has placed her desired object and that it has been moved in her absence, before being asked, "Where will she look for it when she returns?" This objection misses the point of the proposal, however. The idea is that infrequent talk about cognitive states combined with indirect-assertion uses of

terms like “think” and “know” lowers the prior probability of the hypothesis that cognitive states are relevant to the topic of conversation. This remains true even in conversations where those terms are not themselves used.

5. Pragmatic reasoning in FB

With the components of our pragmatic account now outlined, consider the version of false-belief task used in the mindreading scale, drawn from Wellman & Liu (2004):

> Children see a toy figure of a boy, together with a sheet of paper with a backpack and a closet drawn on it. The experimenter says, “Here’s Scott. Scott wants to find his mittens. His mittens might be in his backpack or they might be in the closet. Really, Scott’s mittens are in his backpack. But Scott thinks his mittens are in the closet.” – “So, where will Scott look for his mittens? In his backpack or in the closet?” (the target question) – “Where are Scott's mittens really? In his backpack or in the closet?” (the reality question). To be correct the child must answer the target question “closet” and the reality question “backpack.”

We suggest that children will likely entertain three main hypotheses about the intent behind the target question. One is the helpfulness-interpretation: *she wants me to help Scott find his mittens*. A second is a knowledge-exhibiting interpretation whose topic is the world (rather than Scott’s psychology): *she wants me to show that I know where the mittens really are*. And the third is the intended psychological-knowledge-exhibiting interpretation: *she wants me to show that I know that Scott will look for his mittens where he thinks they are*. On the one hand, the syntax of the question favors

the third hypothesis. But the child is alert for opportunities to be helpful, and will be aware that Scott won't find his mittens unless he looks in the right place. And in addition, as we pointed out above, people's cognitive states are rarely directly relevant to the topic of conversation in the child's previous experience. These factors may render the intended interpretation the least plausible of the three. And even if they don't, since the other two alternatives motivate the same reality-oriented answer, when combined they may lead the child to answer accordingly.

While our account can explain why children fail change-of-location false-belief tasks, it might seem that it is less well placed to explain failures in unexpected-contents versions of the task. For in such cases there is no overt goal. The child is merely asked what someone else (or her previous self) will think is in the Smarties tube (having just discovered for herself that the tube contains pencils and not candies).[44] And in the self-directed version of the task, especially, it may seem unlikely to the child that she is being invited to offer help to her own past self when she is asked, "What *did* you think was in there?" A world-directed knowledge-exhibiting interpretation will be especially salient in unexpected-contents forms of false-belief task, however. For consider what has taken place from the child's perspective. She begins by presuming likely pedagogic intent following initial eye-contact with the experimenter and/or the use of child-directed speech. She is then shown something surprising about a Smarties-tube (that it contains pencils). She might reasonably infer that this is what she is supposed to have learned, and thus exhibit her knowledge of the actual contents of the container when asked.

[44] Nevertheless, a goal may often be tacitly presumed. The child might reasonably assume that everyone likes smarties, and hence be motivated to prevent the target agent from being disappointed to discover that the box contains pencils.

Moreover, recall that "think" is generally used in statements to indirectly assert the complement clause. In second-person questions, likewise, the topic is the complement clause: if you ask me, "Do you think it will rain this afternoon?", the most likely situation is one in which you are asking me about the weather, not my beliefs. This may combine with the saliency of the surprising fact the child has just learned to make it seem likely that she is being asked to exhibit her knowledge of the contents of the container, rather than her own prior beliefs.

Notice that our proposed account of young children's failures in verbal false-belief tasks comports quite nicely with many of the known predictors of false-belief performance. It makes sense, for example, that both executive function and general verbal ability should correlate with false-belief performance. For in order to pass, a child needs not only to decipher the experimenter's query correctly, but also to inhibit answers suggested by alternative interpretations. Likewise, one might expect that verbal ability would depend partly on greater conversational experience, leading to an appreciation that the syntax of the question ("Where will she look?" or "What does she think?") increases the likelihood that the questioner's pragmatic intent has to do with the cognitive states of the story protagonist.

It also makes good sense that false-belief performance should correlate with the extent to which mental-state terms are used in the child's home and with the frequency of child-directed questions (Howard, Mayeux, & Naigles, 2008), and that it should likewise correlate with the number of the child's siblings (McAlister & Peterson, 2013; Perner et al., 1994)—at least, on the assumption that multiplying perspectives in the home is likely to lead to more talk about mental states. The same

point holds for the finding that deaf children of hearing parents, who are delayed in their exposure to language, should also be delayed in verbal false-belief performance (Peterson & Wellman, 2009), and that it should be increased exposure to mental-state terms in particular that predicts subsequent success (Pyers & Senghas, 2009).

In addition, our pragmatic proposal can explain why some variations in task-parameters can reliably shift the age of passing false-belief tasks forward by a few months (Wellman et al. 2001). One factor is whether the target object remains present when the child is asked the test question, or has been removed from the scene. This is generally explained in terms of reduced demands on executive function. And this may well be partly correct. But it can also be explained in pragmatic terms. For if the true location of the target object is unknown to the child, then that will lower the probability that the experimenter is inviting the child to be helpful to the agent in the story, as well as lowering the probability that she is being asked to display knowledge of the actual location; and it will correspondingly increase the probability that the experimenter is inviting the child to display her knowledge of the protagonist's psychology.

If the context is a deceptive one, too, then it is correspondingly less likely that the adult is inviting the child to be helpful. Of course this isn't ruled out. Sometimes tricks are intended just to elicit surprise ("Look where your ball is now!") rather than consternation ("My ball has gone!"). But given an intent to deceive the target agent about the location or nature of an object, it is significantly less likely that the adult will at the same time invite the child to *undeceive* (be helpful to) the agent. A similar point holds if the child herself participates in the experimental transformation. If the

child was encouraged to put the pencils in the Smarties container, then that should lower the likelihood that the adult is now asking the child to be helpful in informing the target agent of this fact. For why would she encourage the child to make the change and then invite the child to helpfully inform the agent of the result? Indeed, this factor may merge with the previous one. That the child is asked to make the move suggests some sort of deceit, or game of hide-and-seek. And then telling about it would spoil the game.

Finally, it also makes sense, of course, that making the protagonist's cognitive states more salient during the setup of a false-belief task should make it easier for children to pass, as we noted earlier (Wellman et al., 2001). For this will help them to see that such states are directly relevant to the topic of conversation, hence raising the likelihood of the psychological-knowledge-exhibiting interpretation.

Our approach also provides an alternative explanation for some recent results that have been thought to support an executive-function account of children's difficulties with standard false-belief tasks. Thus Scott et al. show in an expectancy-violation looking-paradigm that 2.5-year-olds can pass when they passively watch an adult participating in a verbal false-belief task (Scott, He, Baillargeon, & Cummins, 2012). They look longer when the adult gives the incorrect (reality-based) answer. But since the infant is just an observer in these circumstances, there will be no pedagogic cues; and one might expect that the helpfulness-interpretation would be less salient because the infant herself is not involved in the task, and has no opportunity to help. Likewise, He et al. (2012) show using anticipatory looking that 2.5-year-olds pass a false-belief task when the question, "I wonder where she will

look for her scissors?" is self-addressed by the experimenter while gazing at the ceiling, while they fail when the same words are directed at them. For there is a pedagogic cue (eye-contact) in the latter case but not the former; and only in the former it is plain that the child is not being invited to help.

One might wonder whether the finding that 3-year-olds can pass a verbal false-belief task when prompted with the question, "What happens next?" (Rubio-Fernández & Geurts, 2013) presents a problem for our account. For why shouldn't children be motivated to help the Duplo character in the narrative, just as they are in a regular change-of-location false-belief task? This would lead them to guide the Duplo character to the actual location of the bananas she wants, rather than the believed location (which is what they actually do). But in fact the experimental procedure makes clear to participants that they are being invited to continue the story. They are invited to pick up the Duplo character and act-out the conclusion. This should induce the child to access her model of the character's psychology, adopting it as her own in pretend-mode, and then acting-out what one should do when occupying that perspective. (One should go to the empty location to retrieve the bananas, of course, because that is where one thinks they are.) In effect, the procedure lowers the probability of a helpfulness-interpretation of the question, substituting in its place an invitation to pretend to *be* the character in the narrative (Westra, 2016a).

We propose, then, that our pragmatic account can offer a well-motivated alternative explanation of the difficulties young children have with verbal false-belief tasks. We suggest, in fact, that the explanation is no less plausible than that offered by constructivists about mindreading, who think that children's failures manifest a

conceptual deficit. In what follows we will apply our framework (together with other factors) to show that one can similarly explain the order of difficulty of the main components of the mindreading scale.

6. Why DB is easier than FB

With our account of the false-belief task in place, we are now in a position to explain children's performance on the other items on the mindreading scale. We begin with why diverse-belief tasks should be easier for young children than false-belief tasks. Here is a description of a diverse-belief task, drawn from Wellman & Liu (2004), to be compared with the description of the false-belief task given in Section 5.

> Children see a toy figure of a girl, together with a sheet of paper with bushes and a garage drawn on it. The experimenter says, "Here's Linda. Linda wants to find her cat. Her cat might be hiding in the bushes or it might be hiding in the garage. Where do you think the cat is? In the bushes or in the garage?" This is the own-belief question. If the child chooses the bushes: "Well, that's a good idea, but Linda thinks her cat is in the garage. She thinks her cat is in the garage." (Or, if the child chooses the garage, she is told Linda thinks her cat is in the bushes.) Then the child is asked the target question: "So where will Linda look for her cat? In the bushes or in the garage?" To be correct the child must answer the target question opposite from her answer to the own-belief question.

In contrast with the false-belief task (where the child is told where the target object really is), in the diverse-belief task the child is initially asked what *she* thinks. From

the child's perspective this would normally be interpreted as a question about the world, rather than about her beliefs as such. It therefore seems likely (given the pragmatic framework articulated in Section 4) that the child will interpret the experimenter's subsequent assertion, "Linda thinks the cat is in the garage" as also an implicit statement about where the cat really is. In effect, she takes the experimenter to be offering evidence of where the cat is actually located. As a result, she is likely to prioritize the experimenter's belief over her own guess, and thus forms the belief that the cat is in the garage.

Now, consider the hypotheses that the child entertains when interpreting the experimenter's query in this task. The intended interpretation will be, *she wants me to show that I know that Linda will look for her cat where she thinks it is.* The worldly-knowledge-exhibiting interpretation will be, *she wants me to show that I know where the cat is*. The helpfulness interpretation will be, *she wants me to help Linda find her cat*. We think it likely that in this task, as in the false-belief task, younger children will favor one of the alternative hypotheses over the intended one. But in contrast to the false-belief task, this misunderstanding makes no difference to children's performance. Since the child has inferred, based on the experimenter's indirect speech act, that the cat is in the garage, all three interpretations will issue in the same answer, namely, the garage. Thus in contrast to the false-belief task, where the differing interpretations yield different responses, here all three interpretations yield the same (correct) response.

If this account is accurate, then the only way for a child to *fail* a diverse-belief task (besides mere confusion, which is more likely in younger children, of course) is

if she ignores or fails to pick up on the indirect assertion of the experimenter, and goes on believing her own guess. Believing that the cat is in the bushes, the helpful thing to tell Linda is that this is where the cat is, which would then get scored as incorrect.[45] (The worldly-knowledge-exhibiting interpretation, in contrast, will seem implausible in this case. For it was not the experimenter who taught her the location of the cat.)

7. Why DD is easier than DB

Having explained from a nativist perspective why the diverse-belief task may be pragmatically easier than the false-belief task, we now turn to explain why the diverse-desire task should be easier still. Here is a canonical description of the task, drawn from Wellman & Liu (2004).

> Children see a toy figure of an adult, together with a sheet of paper with a carrot and a cookie drawn on it. The experimenter says, "Here is Mr. Jones. It's snack time, so Mr. Jones wants a snack to eat. Here are two different snacks: a carrot and a cookie. Which snack would you like best? Would you like a carrot or a cookie best?" This is the own-desire question. If the child chooses the carrot: "Well, that's a good choice, but Mr. Jones really likes cookies. He doesn't like carrots. What he likes best are cookies." (Or, if the child chooses the cookie, she is told that Mr. Jones likes carrots.) Then the child is asked the target question: "So, now it's time to eat. Mr. Jones can only

[45] Why would children ever *believe* a mere guess? One possibility is that children (and especially young children) are chronically poor at source monitoring (Bruck & Ceci, 1999). Having guessed an answer (or indeed, having merely been asked to imagine a particular state of affairs), children are apt thereafter to speak and behave as if they really believe it.

choose one snack, just one. Which snack will Mr. Jones choose? A carrot or a cookie?" To be scored as correct, the child must answer the target question opposite from her answer to the own-desire question.

This task is pragmatically easy for children for two related reasons. The first is that we know from corpus-data that young children have plenty of experience with conversations in which talk of people's desires takes place, and are the topic of conversation, as well as with conversations in which there is encouragement for children to tell others what they want (MacWhinney, 2014; see also Tamoepeau & Ruffman, 2006).

Further, while "think" is most often used in an indirect manner, the same is unlikely to be true of "want." Even though "want" *can* be used as an indirect way of communicating an imperative (e.g. "Do you want to be quiet?" can mean, "Be quiet!"), our suspicion is that such uses are comparatively rare. In fact children's desires, unlike their other thoughts, are frequent topics of conversation in child-directed speech: primary caregivers continuously monitor and manage their children's needs, so it makes sense to ask them what they want on a regular basis. Children's beliefs, in contrast, are less vital to the caregiving process. Thus, desire-talk in general, and the verb "want" in particular, is much less likely to pose pragmatic challenges for a novice speaker than does talk of cognitive states like belief and knowledge. Hence the conversation initiated with the child in a diverse-desire task is comparatively unambiguous. Most young children can easily figure out that they are being told what Mr. Jones wants, and can predict what he will choose accordingly.

In fact the only way a child can fail the diverse-desire task (aside from being

completely confused, mishearing the question, and so on) is if she thinks that everybody shares her desires, and refuses to accept the statement that Mr. Jones likes carrots. If she believes that Mr. Jones (like everyone) prefers cookies to carrots, then she will answer, "Cookies" when asked what Mr. Jones will choose. There is some evidence that children might reason egocentrically like this at an early age. Repacholi and Gopnik show this pattern of response at 14 months, but not 18 months (Repacholi & Gopnik, 1997). However, although children are fairly adept at giving and helping by this age (Warneken & Tomasello, 2007), the test question in this experiment was pragmatically demanding. In the pre-test phase, the experimenter expressed a preference for either broccoli or crackers. Then in the test phase, bowls of broccoli and crackers were placed between the child and the experimenter. The experimenter then asked, "Can you give me some?" and the child had to interpret the nature of the request. ("Some *what*?" she might wonder.) When 14-month-olds failed, this was interpreted as egocentrism. However, it could just be that they found the unusual nature of the request confusing—especially when, for them, the crackers are the most salient option (Baillargeon et al., 2015). In fact, even at much earlier ages children already show an understanding of goals and preferences, and are surprised when agents act contrary to their preferences (e.g. Woodward, 1998; Luo & Baillargeon, 2005). We argue, then, that there is no age at which infants are truly egocentric about desires.

Indeed, we predict that it might be easier still for young children to pass a version of the diverse-desire test if it were to involve two protagonists, especially if the desired items were affectively-neutral for the child (or even better: novel).

Children could be introduced to Mr. Jones and Mr. Smith. They are told that Jones likes daxes, whereas Smith likes blickets, while being shown some of each. They could then be asked: "Which will Mr. Jones choose?" Here there would be no opportunity for the child's response to be biased by her own preferences.

8. Knowledge-Access

We turn now to the knowledge-access task, and the way in which performance on the task is influenced by cultural background. Here is a canonical statement of the task, drawn from Wellman & Liu (2004).

> Children see a nondescript plastic box with a closed drawer (which contains a small plastic toy dog inside). The experimenter says, "Here's a drawer. What do you think is inside the drawer?" (The child can give any answer she likes, or indicate that she does not know). Next, the drawer is opened and the child is shown the contents of the drawer. The experimenter says, "Let's see ... it's really a dog inside!" The drawer is then closed: "Okay, what is in the drawer?" Then a toy figure of a girl is produced: "Polly has never ever seen inside this drawer. Now here comes Polly. So, does Polly know what is in the drawer?" (the target question) "Did Polly see inside this drawer?" (the memory question). To be correct the child must answer the target question "no" and the memory control question "no."

Why is this task easier than the false-belief task? In contrast with the latter, there is nothing in the task to suggest that Polly has the goal of finding the toy dog. So an interpretation of the test question as inviting the child to be helpful to Polly is

correspondingly less likely. One of the main factors that pushes children toward incorrect answers in a false-belief task is therefore absent. However, children might still assume that the most salient fact they have learned about the situation is that the drawer contains a toy dog. They might therefore hear the test question as an invitation to show what they have learned. Moreover, the intended-interpretation of the test question still makes cognitive states a topic of conversation, which is something that children at this age would regard as unusual. So there are still factors (albeit fewer factors) biasing young children toward an incorrect answer, thus explaining why the knowledge-access task is harder than the diverse-desire task.

Is this the only reason why the knowledge-access task is harder than the diverse-desire task (and, for Western subjects, the diverse-belief task)? In fact there is good reason to think that at least part of the difficulty results from an experimental artifact. For the knowledge-access task (alone among tests in the mindreading scale) requires children to give yes/no answers.[46] (Both answers need to be negative for the child to be scored as correct.) Yet we know that children of this age are strongly biased to answer all yes/no questions positively (Fritzley & Lee, 2003; Okanda & Itakura, 2008). The yes-bias is especially strong in younger children, but begins to weaken through the fourth year of life (at about the time children begin passing the knowledge-access task). So one reason why young children pass the diverse-desire task before they pass the knowledge-access task is likely to be this: the latter, but not the former, involves yes/no questions that require a negative answer. Moreover, a general capacity to inhibit the yes-bias is strongly predicted by both verbal ability and

[46] The same form of yes/no question was also used in the studies cited in Wellman & Liu (2004), on which the knowledge-access task was based (Fabricius & Khalil, 2003; Flavell, Flavell, Green, & Moses, 1990; Surian & Leslie, 1999).

executive (specifically inhibitory) control, even after controlling for age (Moriguchi, Okanda, & Itakura, 2008), both of which will be developing during this period.

If these explanations of the difficulty of the knowledge-access task in comparison to the diverse-desire and false-belief tasks are correct, then that leaves us with the question of cross-cultural differences in the relative ease of the diverse-belief and knowledge-access tasks. Why do children in countries like America and Australia find the former easier than the latter, whereas in countries like China and Iran the reverse is true? One part of the explanation turns on the role of the yes-bias in the knowledge-access task. For we know that children in "collectivist" cultures tend to perform better on measures of inhibitory control (Lan, Legare, Ponitz, Li, & Morrison, 2011; Oh & Lewis, 2008; Sabbagh et al., 2006), and that inhibitory control is strongly predictive of children's ability to overcome the yes-bias (Moriguchi et al., 2008). So one component of the explanation is simply that children from such cultures are able to overcome the yes-bias earlier than do children from "individualist" ones.

In addition, we can also adopt Wellman's own proposal, but giving it a pragmatic rather than a constructivist twist. He points out that in "collectivist" cultures there is much greater emphasis on the importance of knowledge, the importance of respecting those who have knowledge, and so on (Wellman et al., 2006; Shahaien et al. 2011). As a result, we suggest, children in these cultures will be better positioned to recognize the conversational importance of the "seeing leads to knowing" principle that is at stake in the knowledge-access task. This raises the likelihood that the child will interpret the experimenter's question as asking about

Polly's state of knowledge rather than the location of the dog, leading her to answer correctly.

Why should there be no difference between "collectivist" and "individualist" cultures in the age at which children pass verbal false-belief tasks, however (Wellman et al., 2001), given that the former have more advanced executive function abilities, and given that executive function predicts some of the individual variance in false-belief, as we noted earlier? Here, too, we can appeal to Wellman's own proposals for an explanation. While the increased executive-function abilities of "collectivist" children should give them a boost in false-belief tasks, at the same time the greater emphasis placed on differences of opinion and conflicts of belief distinctive of "individualist" cultures should provide additional help to Western children in figuring out the intent behind the false-belief question. So the two factors may cancel one another out.

9. The benefits of training

We have suggested explanations of the relative ordering of the main components of the mindreading-scale, and have sketched an account of cultural variation in the scale's intermediate components. We now turn to consider the impact of training on false-belief performance.

Rhodes & Wellman (2013) undertook a training study with American children, who normally find diverse-belief tasks significantly easier (and pass them 3–6 months earlier) than knowledge-access tasks. They found that children who are already capable of passing both forms of task, while still failing false-belief tasks, are

more likely to benefit from training in the latter. Significantly more of these children transition from failing to passing false-belief tasks following training. Children who are only capable of passing diverse-belief tasks, in contrast (who prior to training fail both knowledge-access and false-belief), are less likely to benefit. Wellman (2014) argues that this finding supports a constructivist position. Children who are further along the mindreading scale are on the cusp of genuinely understanding the nature of false beliefs, and can benefit from training that targets that understanding, whereas those earlier along the scale are not, and do not.

From our nativist perspective, in contrast, the difference between diverse-belief tasks and false-belief tasks is one of pragmatic difficulty, not conceptual understanding. (See Sections 5 and 6.) In particular, the same pragmatic misunderstanding that leads children to fail a false-belief task (interpreting the question in one or other reality-oriented way) will lead children to answer in a manner that is scored as passing a diverse-belief task. But it normally takes 6–12 months for children to transition from passing the diverse-belief task to passing the false-belief task. During this time they have more and more experience, both of conversations in which cognitive states are the main topic of conversation and of questions that are intended to elicit statements of what they know, and they learn to discriminate some of the cues that will enable them to tell when these things are so. For instance, children will start to see that facts about beliefs become especially noteworthy and relevant in false-belief scenarios, and that explicit talk about beliefs is associated with particular syntactic frames, such as complementation syntax (Papafragou et al. 2007). They will also start to realize that we often make implicit reference to beliefs. In

addition, children will develop more general pragmatic competence during this period. This will help them to develop more refined expectations about how various question-types (such as knowledge-exhibiting questions and invitations to be helpful) tend to be asked. All of these experiences will lead them to gradually update their priors about the conversational relevance of cognitive states, and help them to successfully apply their knowledge of beliefs in conversation.

In contrast, children who have recently become capable of passing the knowledge-access task are only 3–6 months away from being able to pass the false-belief task (Rhodes & Wellman, 2013). It makes sense, then, that they will already have accumulated much of the necessary conversational experience and sensitivity. As a result, they are better positioned to respond to training in false-belief tasks. They improve (whereas children who fail knowledge-access tasks do not), not because the training induces a conceptual understanding of false belief, but simply because the training makes talk of mental states more salient to them—thereby increasing the prior probability of the interpretation intended by the experimenter. In children who do not yet pass knowledge-access tasks, in contrast, the lure of the helpfulness and real-world knowledge-exhibiting interpretations is still too strong for them to show any improvement.

At this point, a defender of the conceptual development account might point out that, on her view, the benefits of training should generalize to a wide range of situations. The conceptual-change account should predict that the benefits of training will extend to mindreading tasks employing quite different questions and materials, involving different experimenters, and taking place in an entirely different context.

Our view, in contrast, might predict that the benefits of training will be quite local and context-specific, and not generalize to new testing situations. What is learned in just a few training episodes (as opposed to months of communicative experience) is more likely to be that in conversations with *these* people taking place in *this* context cognitive states are relevant to the interpretation of *these sorts of questions*.[47] While these predictions have not been directly tested, there are some results in the literature that might seem to support the conceptual-change account. We will discuss these briefly here.

Lecce et al. show that training both children and aging adults on false-belief tasks not only improves false-belief performance, but also benefits later performance on metamemory tasks, enabling them to exhibit greater knowledge of such facts as that it is easier to learn a short list than a long one, or that it is easier to learn in the absence of distractions (Lecce, Bianco, Demicheli, & Cavallini, 2014; Lecce, Bottiroli, Bianco, Rosi, & Cavallini, 2015). But the transfer, here, is quite unlikely to be conceptual. For how could training on false-belief tasks lead people to acquire explicit knowledge of these sorts, and to do so in just a couple of days? Rather, we suggest that the training increases the saliency of mental-state talk in general. Consistent with this interpretation, Lecce et al. (2015) found that false-belief training significantly *decreases* performance on physical-causality tasks. It is surely more likely that false-belief training decreases the saliency of physical-causal talk than that it causes some sort of physical-causality forgetting, or results in some kind of conceptual loss.

[47] This will depend on the nature of the training, of course. If the training takes place across many different contexts, then it may succeed in raising the probability that cognitive states are a topic of conversation across the board.

In addition, there is the recent finding that false-belief training helps children in a game of deceive-the-experimenter, where to win they have to tell the experimenter the opposite of what they know to be true (Ding, Wellman, Wang, Fu, & Lee, 2015). But this, too, is explicable in terms of a kind of pragmatic saliency-priming. For repeated exposure to false-belief tasks should make other people's cognitive states more salient when engaging in verbal interactions with the experimenters, making it easier for them to adopt the false-belief-causing response. And it should also make clear to children that a helpfulness response (telling the experimenter where the target really is, thereby enabling him to win) is not what is being looked for in the interaction (any more than it is in false-belief tasks).

10. Conclusion

Our goal in this paper has been to show that the mindreading-scale data do not presently support constructivism over nativism. This is because there are plausible and empirically well-motivated alternative explanations of those results that are consistent with nativism about mindreading, mostly of a pragmatic sort. Our main proposal has been that it takes children a while to figure out that cognitive states can be a topic of conversation and to develop the pragmatic skills to discern when this is so. They have to learn that sometimes questions are really invitations for them to display their psychological knowledge rather than requests to be helpful or to display their knowledge of the worldly facts. And we have suggested that the various components of the mindreading scale differ in their pragmatic demands rather than their conceptual difficulty. But we have, of course, done nothing to *support* the truth of nativism here. That is a task for another occasion. Our goal has merely been to

show that the robustness of the mindreading scale (and children's late-emerging performance in verbal false-belief tasks, in particular) provides inadequate reason to *reject* a nativist account.

Nevertheless, our pragmatic account is surely ripe for experimental testing. For it makes numerous predictions for interventions that should impact children's performance in these tasks which would distinguish it from constructivist approaches. Here we will mention just one. Manipulations that draw children's attention to cognitive states as the topic of conversation should improve performance. For instance, one should be able to "prime" children who are on the cusp of passing false-belief tasks into succeeding, by engaging them in conversations in which cognitive states of belief or knowledge are the topic. Whether those conversations concern false beliefs in particular, or provide feedback that could be construed as evidence for a representational theory of mind, should be irrelevant. The same effect could also be achieved through task designs that highlight the salience of cognitive states—for instance, by making them highly relevant to children's goals (e.g. Dudley et al., 2015).

Bibliography

Ames, D. L., & Fiske, S. T. (2013). Outcome dependency alterns the neural substrates of impression formation. *NeuroImage*, *83*, 599–608.

Ames, D. L., Fiske, S. T., & Todorov, A. (2011). *Impression formation: A focus on others' intents*. *The Oxford handbook of social neuroscience*. Oxford: Oxford University Press.

Ames, D. R., Flynn, F. J., & Weber, E. U. (2004). It's the thought that counts: on perceiving how helpers decide to lend a hand. *Personality and Social Psychology Bulletin*, *30*(4), 461–474.

Amsterlaw, J., & Wellman, H. M. (2006). Theories of Mind in Transition: A Microgenetic Study of the Development of False Belief Understanding. *Journal of Cognition and Development*, *7*(2), 139–172.

Andrews, K. (2008). It's in your nature: A pluralistic folk psychology. *Synthese*, *165*(1), 13–29.

Andrews, K. (2012). *Do apes read minds?: Toward a new folk psychology*. Cambridge, MA: MIT Press.

Anscombe, G. E. M. (1958). Modern moral philosophy. *Philosophy*, *33*(124), 1–19.

Apperly, I. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind."* Psychology Press.

Apperly, I. (2013). Can theory of mind grow up? Mindreading in adults, and its implications for the development and neuroscience of mindreading. In S. Baron-Cohen, H. Tager-Flusberg, & M. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (3rd ed., pp. 72–92).

Oxford: Oxford University Press.

Apperly, I., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970.

Avramides, A. (2000). *Other minds*. New York, NY: Routledge.

Baillargeon, R. (2008). Innate ideas revisited for a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, *3*(1), 2–13.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–8.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278.

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., … Laurence, S. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1755), 20122654.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological Review*, *113*(3), 628–47.

Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental Science*, *8*(6), 492–499.

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science (New York, N.Y.)*, *324*(5931), 1160–4.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences*, *1*(04), 557.

Benson, J. E., & Sabbagh, M. A. (2005). Theory of Mind and Executive Functioning: A Developmental Neuropsychological Approach. In P. D. Zelazo, M. Chandler, & E. Crone (Eds.), *The Developing Infant Mind: Integrating Biology and Experience* (pp. 63–80). New York, NY: Guilford Press.

Bermudez, J. L. (2003). The Domain of Folk Psychology. *Royal Institute of Philosophy Supplement*, 25–48.

Bloom, L., Rispoli, M., Gartner, B., & Hafitz, J. (1989). Acquisition of complementation. *Journal of Child Language*, *16*(01), 101.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). *Investigating Action Understanding: Inferential Processes versus Action Simulation*. *Current Biology* (Vol. 17).

Bratman, M. (1992). Shared Cooperative Activity. *Philosophical Review*, *101*(2),

327–341.

Bräuer, J., Call, J., & Tomasello, M. (2004). Visual perspective taking in dogs (Canis familiaris) in the presence of barriers. *Applied Animal Behaviour Science*, *88*(3-4), 299–317.

Bruck, M., & Ceci, S. J. (1999). The suggestibility of children's memory. *Annual Review of Psychology*, *50*(1), 419–439.

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications*, *7*, 10506.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337–342.

Buttelmann, D., Over, H., Carpenter, M., & Tomasello, M. (2014). Eighteen-month-olds understand false beliefs in an unexpected-contents task. *Journal of Experimental Child Psychology*, *119*, 120–6.

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

Butterfill, S., & Apperly, I. (2013). How to Construct a Minimal Theory of Mind. *Mind and Language*, *28*(5), 606–637.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Development*, *70*(2), 381–95.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30

years later. *Trends in Cognitive Sciences*, *12*(5), 187–92.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development*, *72*(4), 1032–1053.

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, *11*(2), 73–92.

Carruthers, P. (2006). *The Architecture of Mind*. Oxford: Oxford University Press.

Carruthers, P. (2013). Mindreading in Infancy. *Mind & Language*, *28*(2), 141–172.

Carruthers, P. (2015a). Mindreading in adults: evaluating two-systems views. *Synthese*, *192*, 1–16.

Carruthers, P. (2015b). *The centered mind: what the science of working memory shows us about the nature of human thought*. Oxford University Press.

Carruthers, P. (2015c). Two Systems for Mindreading? *Review of Philosophy and Psychology*, *6*.

Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*. Cambridge University Press.

Cassam, Q. (2007). *The Possibility of Knowledge*. Oxford: Oxford University Press.

Chandler, M., Fritz, A. S., Hala, S., Chandler, M., Fritz, A. S., & Hala, S. (1989). Articles Small-Scale Deceit : Deception as a Marker of Theories of Mind. *Child Development*, *60*(6), 1263–1277.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*.

Chaumon, M., Kveraga, K., Barrett, L. F., & Bar, M. (2014). Visual predictions in the orbitofrontal cortex rely on associative content. *Cerebral Cortex*, *24*(11), 2899–907.

Chen, Y., & Zhao, Y. (2015). Intergroup threat gates social attention in humans.

Chisholm, R. (1967). Brentano on descriptive psychology and the intentional. In E. N. Lee & M. Mandelbaum (Eds.), *Phenomenology and Existentialism*.

Choi, I., & Nisbett, R. E. (1998). Situational salience and cultural differences in the correspondence bias and actor-observer bias. *Personality and Social Psychology Bulletin*.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*(1), 47–63.

Christensen, W., & Michael, J. (2015). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*(A), 48–64.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*(05), 461–480.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395.

Cloutier, J., Gabrieli, J. D. E., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: When people are not who we expect them to be. *NeuroImage*.

Cogsdill, E. J., Todorov, A., Spelke, E. S., & Banaji, M. R. (2014). Inferring Character From Faces: A Developmental Study. *Psychological Science*, *25*(5), 1132–1139.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences*, *3*(3), 115–120.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, *58*(3), 306–24.

Csibra, G. (2008). Action mirroring and action understanding: an alternative account. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorymotor Foundations of Higher Cognition. Attention and Performance XXII* (pp. 435–459). Oxford: Oxford University Press.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–53.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631–48.

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1–33.

Cunningham, W. a., Zelazo, P. D., Packer, D. J., & Van Bavel, J. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, *25*(5), 736–760.

d'Souza, D., & Karmiloff-Smith, A. (2011). When modularization fails to occur: a developmental perspective. *Cognitive Neuropsychology*, *28*(3-4), 276–287.

Dalmaso, M., Pavan, G., Castelli, L., & Galfano, G. (2012). Social status gates social attention in humans. *Biology Letters*, *8*(3), 450–452.

de Bruin, L., & Strijbos, D. (2015). Direct social perception, mindreading and Bayesian predictive coding. *Consciousness and Cognition*, *36*, 565–570.

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507.

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition : a longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, *17*, 1037–1060.

Deen, B., & Saxe, R. R. (2012). Neural correlates of social perception: The posterior superior temporal sulcus is modulated by action rationality, but not animacy. In *Proceedings of the 33rd Annual Cognitive Science Society Conference* (pp. 276–281).

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, *1*(04), 568.

Devine, R. T., & Hughes, C. (2014). Relations Between False Belief Understanding and Executive Function in Early Childhood: A Meta-Analysis. *Child Development*, *85*(5), 1777–1794.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge. Evidence from 9-month-old infants. *Psychological Science*, *21*(12), 1871–7.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics*, *12*(2), 97–141.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-Mind Training Causes Honest Young Children to Lie. *Psychological Science*, *26*(11), 1812–1821.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge, UK: Cambridge University Press.

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-olds' understanding of know and think. In F. Schwarz (Ed.), *Experimental Perspectives on Presuppositions* (pp. 241–262). Springer.

Duh, S., Paik, J. H., Miller, P. H., Gluck, S. C., Li, H., & Himelfarb, I. (2016). Theory of mind and executive function in Chinese preschool children. *Developmental Psychology*, *52*(4), 582–591.

Dunn, J., & Brophy, M. (2005). Communication, Relationships, and Individual Differences in Children's Understanding of Mind. In J. W. Astington & J. A. Baird (Eds.), *Why Language Matters for Theory of Mind* (pp. 50–69). Oxford: Oxford University Press.

Elekes, F., Varga, M., & Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, *41*, 93–103.

Fabricius, W. V., & Khalil, S. L. (2003). False Beliefs or False Positives? Limits on Children's Understanding of Mental Representation. *Journal of Cognition and Development*, *4*(3), 239–262.

Farroni, T., Massaccesi, S., Pividori, D., & Johnson, M. H. (2009). Gaze Following in

Newborns. *Infancy*, *5*(1), 39–60.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology*, *70*(6), 1164.

Fenici, M. (2013). Social cognitive abilities in infancy: Is mindreading the best explanation? *Philosophical Psychology*, (September 2014), 1–25.

Ferrari, C., Vecchi, T., Todorov, A., & Cattaneo, Z. (2016). Interfering with activity in the dorsomedial prefrontal cortex via TMS affects social impressions updating. *Cognitive, Affective, & Behavioral Neuroscience*, 626–634.

Fiebich, A., & Coltheart, M. (2015). Various Ways to Understand Other Minds: Towards a Pluralistic Approach to the Explanation of Social Understanding. *Mind and Language*, *30*(3), 235–258.

Fiske, S. T. (2015). Intergroup biases: A focus on stereotype content. *Current Opinion in Behavioral Sciences*, *3*(April), 45–50.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, *17*(1), 99–103.

Flavell, J. H., Flavell, E. R., Green, F. L., & Moses, L. J. (1990). Young Children's Understanding of Fact Beliefs versus Value Beliefs. *Child Development*, *61*(4), 915–928.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition*, *44*(3), 283–296.

Foot, P. (1967). *Theories of ethics*. Oxford: Oxford University Press.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490–495.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521).

Fritzley, V. H., & Lee, K. (2003). Do young children always say yes to yes-no questions? A metadevelopmental study of the affirmation bias. *Child Development*, *74*(5), 1297–1313.

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*, *17*(2), 535–543.

Gallagher, S., & Povinelli, D. J. (2012). Enactive and Behavioral Abstraction Accounts of Social Understanding in Chimpanzees, Infants, and Adults. *Review of Philosophy and Psychology*, *3*(1), 145–169.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–409.

Gergely, G. (2013). Ostensive Communication and Cultural Learning: The Natural Pedagogy Hypothesis. *Agency and Joint Attention*, 139.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gilbert, D. T., Malone, P. S., Aronson, J., Giesler, B., Higgins, T., Ross, L., … Trope, Y. (1995). The Correspondence Bias. *Psychological Bulletin*, *117*(1), 21–38.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733–740.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Goodman, N. (1955). *Fact, Fiction, & Forecast*. Cambridge, MA: Harvard University Press.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, *59*(1), 26–37.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, *8*(8), 371–7.

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, *7*(1-2), 145–171.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–108.

Gordon, R. M. (1986). Folk Psychology as Simulation. *Mind & Language*, *1*(2), 158–171.

Gray, C. (2007). *Writing social stories with Carol Gray*. Future Horizons.

Grice, H. P. (1991). *Studies in the Way of Words*. Harvard University Press.

Hacquard, V. (2014). Bootstrapping attitudes. *Proceedings of SALT*, (24), 330–352.

Hadwin, J., & Perner, J. (1991). Pleased and surprised: Children's cognitive theory of emotion. *British Journal of Developmental Psychology*, *9*(2), 215–234.

Hala, S., Chandler, M., & Fritz, A. S. (1991). Fledgling Theories of Mind: Deception as a Marker of Three-Year-Olds' Understanding of False Belief. *Child Development*, *62*(1), 83–97.

Hale, C. M., & Tager-Flusberg, H. (2003). The Influence of Language on Theory of Mind: A Training Study. *Developmental Science*, *6*(3), 346–359.

Hansen, M. B. (2010). If You Know Something, Say Something: Young Children's Problem with False Beliefs. *Frontiers in Psychology*, *1*, 23.

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences*, *1*(04), 576.

Harman, G. (1999). Moral Philosophy Meets Social Psychology : Virtue Ethics and the Fundamental Attribution Error Author (s): Gilbert Harman Source :

Proceedings of the Aristotelian Society , New Series , Vol . 99 (19. *Proceedings of the Aristotelian Society, New Series*, *99*, 315–331.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. *NeuroImage*, *28*(4), 763–769.

Harris, P. L., de Rosnay, M., & Pons, F. (2005). Language and Children's Understanding of Mental States. *Current Directions in Psychological Science*, *14*(2), 69–73.

Haslam, N., Bastian, B., & Kashima, Y. (2006). Psychological Essentialism, Implicit Theories, and Intergroup Relations. *Group Processes and Intergroup Relations*, *9*(1), 63–76.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2013). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, *24*(8), 1979–1987.

He, Z., Bolz, M., & Baillargeon, R. (2012). 2.5-year-olds succeed at a verbal anticipatory-looking false-belief task. *British Journal of Developmental Psychology*, *30*(1), 14–29.

Heal, J. (1996). Simulation, theory, and content. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 75–89). Cambridge, UK: Cambridge University Press.

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences*, *18*(4), 167–70.

Helming, K. A., Strickland, B., & Jacob, P. (2016). Solving the Puzzle about Early

Belief-Ascription. *Mind & Language*, *31*(4), 438–469.

Heyes, C. (2014a). False belief in infancy: A Fresh Look. *Developmental Science*, 1–13.

Heyes, C. (2014b). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(2), 131–43.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J., & Palmer, C. (2014). Social Cognition as Causal Inference: Implications for Common Knowledge and Autism. In M. Gallotti & J. Michael (Eds.), *Perspectives on Social Ontology and Social Cognition* (pp. 167–189). Dordrecht: Springer Netherlands.

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's Eyes Trigger Shifts of Visual Attention in Human Infants. *Psychological Science*, *9*(2), 131–134.

Hooper, J. (1975). On assertive predicates. *Syntax and Semantics*, *4*, 91–124.

Hooper, N., Ergogan, A., Keen, G., Lawton, K., & McHugh, L. (2015). Perspective taking reduces the fundamental attribution error.pdf. *Journal of Contextual Behavioral Science*, *4*, 69–72.

Hornstein, N. (2005). Empiricism and rationalism as research strategies. *The Cambridge Companion to Chomsky*, 145.

Howard, A. A., Mayeux, L., & Naigles, L. R. (2008). Conversational correlates of children's acquisition of mental verbs and a theory of mind. *First Language*,

28(4), 375–402.

Hughes, C., Devine, R. T., Ensor, R., Koyasu, M., Mizokawa, A., & Lecce, S. (2014). Lost in Translation? Comparing British, Japanese, and Italian Children's Theory-of-Mind Performance. *Child Development Research*, *2014*, 1–10.

Hume, D. (2000). *An enquiry concerning human understanding: A critical edition*. (T. L. Beauchamp, Ed.) (Vol. 3). Oxford: Oxford University Press.

Hutto, D. D. (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. MIT Press.

Hyslop, A. (2014). Other Minds. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Summer 201).

Hyun, J.-S., & Luck, S. J. (2007). Visual working memory as the substrate for mental rotation. *Psychonomic Bulletin & Review*, *14*(1), 154–158.

Icard, T. (2016). Subjective Probability as Sampling Propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: Cookies and kindness. *Journal of Personality and Social Psychology*, *21*(3), 384–388.

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences*, *9*(1), 21–25.

Jacques, S., & Zelazo, P. D. (2005). Language and the Development of Cognitive Flexibility: Implications for Theory of Mind. In *Why Language Matters for Theory of Mind* (pp. 144–162). Oxford: Oxford University Press.

Jakobsen, K. V., Frick, J. E., & Simpson, E. A. (2013). Look Here! The Development of Attentional Orienting to Symbolic Cues. *Journal of Cognition and*

Development, *14*(2), 229–249.

Jastorff, J., Clavagnier, S., Gergely, G., & Orban, G. A. (2011). Neural Mechanisms of Understanding Rational Actions: Middle Temporal Gyrus Activation by Contextual Violation. *Cerebral Cortex*, *21*(2), 318–329.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objets: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences*, *18*(7), 314–320.

Johnson, C. N., & Maratsos, M. P. (1977). Early Comprehension of Mental Verbs: Think and Know. *Child Development*, *48*(4), 1743–1747.

Jones, E., & Harris, A. (1967). The Attribution of Attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24.

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and Brain Sciences*, *34*(4), 169–88; disuccsion 188–231.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan.

Kalish, C. W. (2002). Children's predictions of consistency in people's actions. *Cognition*, *84*(3), 237–265.

Kampis, D., Parise, E., Csibra, G., & Kovács, Á. M. (2015). Neural signatures for sustaining object representations attributed to others in preverbal human infants. *Proceedings of the Royal Society of London B: Biological Sciences*, *282*(1819).

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kestemont, J., Vandekerckhove, M., Ma, N., Van Hoeck, N., & Van Overwalle, F. (2013). Situation and person attributions under spontaneous and intentional instructions: An fMRI study. *Social Cognitive and Affective Neuroscience*, *8*(5), 481–493.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5), e36399.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166.

Kingstone, A., Tipper, C., Ristic, J., & Ngan, E. (2004). The eyes have it!: An fMRI investigation. *Brain and Cognition*, *55*(2), 269–271.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.

Kitayama, S., Duffy, S., Kawamura, T., & Larsen, J. T. (2003). Perceiving an object and its context in different cultures: A cultural look at new look. *Psychological Science*, *14*(3), 201–206.

Kloo, D., & Perner, J. (2003). Training Transfer Between Card Sorting and False Belief Understanding: Helping Children Apply Conflicting Descriptions. *Child Development*, *74*(6), 1823–1839.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knudsen, B., & Liszkowski, U. (2012). Eighteen- and 24-month-old infants correct

others in anticipation of action mistakes. *Developmental Science*, *15*(1), 113–122.

Knudsen, E. I. (2011). Control from below: the role of a midbrain network in spatial attention. *The European Journal of Neuroscience*, *33*(11), 1961–72.

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron*, *79*(5), 836–848.

Kovács, Á. M. (2015). Belief files in theory of mind reasoning. *Review of Philosophy and Psychology*.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–4.

Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., & Sodian, B. (2006). Skalierung von "Theory of Mind"-Aufgaben. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *38*(4), 186–195.

Krull, D. S., Hui-Min Loy, M., Lin, J., Wang, C.-F., Chen, S., & Zhao, X. (1999). The Fundamental Fundamental Attribution Error: Correspondence Bias in Individualist and Collectivist Cultures.pdf. *Personality & Social Psychology Bulletin*, *25*(10), 1208–1219.

Krull, D. S., Seger, C. R., & Silvera, D. H. (2008). Smile when you say that: Effects of willingness on dispositional inferences. *Journal of Experimental Social Psychology*, *44*(3), 735–742.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos (Ed.), *Criticism and the Growth of Knowledge* (pp. 91–196). Cambridge, UK: Cambridge University Press.

Lan, X., Legare, C. H., Ponitz, C. C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child Psychology*, *108*(3), 677–692.

Lecce, S., Bianco, F., Demicheli, P., & Cavallini, E. (2014). Training Preschoolers on First-Order False Belief Understanding: Transfer on Advanced ToM Skills and Metamemory. *Child Development*, *85*(6), n/a–n/a.

Lecce, S., Bottiroli, S., Bianco, F., Rosi, A., & Cavallini, E. (2015). Training older adults on Theory of Mind (ToM): Transfer on metamemory. *Archives of Gerontology and Geriatrics*, *60*(1), 217–226.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind". *Trends in Cognitive Sciences*, *8*(12), 528–33.

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*(1), 45–85.

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science*, *1*(2), 247–253.

Lewis, D. (1969). *Convention: A philosophical study*. John Wiley & Sons.

Lewis, S. (2013). *Pragmatic enrichment in language processing and development*. University of Maryland. Retrieved from http://drum.lib.umd.edu/handle/1903/14599

Lewis, S., Hacquard, V., & Lidz, J. (2012). The semantics and pragmatics of belief reports in preschoolers. *Proceedings of SALT*, *22*, 247–267.

Li, V., Spitzer, B., & Olson, K. R. (2014). Preschoolers Reduce Inequality While

Favoring Individuals With More. *Child Development*, *85*(3), 1123–1133.

Liepelt, R., & Brass, M. (2010). Top-Down Modulation of Motor Priming by Belief About Animacy, *57*(3), 221–227.

Liepelt, R., & Cramon, D. Y. Von. (2008). What Is Matched in Direct Matching ? Intention Attribution Modulates Motor Priming, *34*(3), 578–591.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age. *Developmental Science*, *10*(2), F1–7.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, *108*(3), 732–739.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. a. (2008). Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology*, *44*(2), 523–31.

Locke, J. (1690). *An essay concerning human understanding*. Dover Publications.

Lohmann, H., & Tomasello, M. (2003). The Role of Language in the Development of False Belief Understanding: A Training Study. *Child Development*, *74*(4), 1130–1144.

Low, J., Drummond, W., Walmsley, A., & Wang, B. (2014). Representing how rabbits quack and competitors act: limits on preschoolers' efficient ability to track perspective. *Child Development*, *85*(4), 1519–34.

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: state of the art. *The British Journal of Developmental Psychology*, *30*(Pt 1), 1–13.

Low, J., & Simpson, S. (2012). Effects of labeling on preschoolers' explicit false belief performance: outcomes of cognitive flexibility or inhibitory control? *Child Development*, *83*(3), 1072–84.

Low, J., & Watts, J. (2013). Attributing false-beliefs about object identity is a signature blindspot in humans' efficient mindreading system. *Psychological Science*, *24*(3), 305–311.

Luo, Y., & Baillargeon, R. (2005). Can a Self-Propelled Box Have a Goal?: Psychological Reasoning in 5-Month-Old Infants. *Psychological Science*, *16*(8), 601–608.

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*, *12*(1), 142–9.

Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F. Van, Seurinck, R., & Fias, W. (2011). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, *7*(8), 937–950.

Ma, N., Vandekerckhove, M., Van Hoeck, N., & Van Overwalle, F. (2012). Distinct recruitment of temporo-parietal junction and medial prefrontal cortex in behavior understanding and trait identification. *Social Neuroscience*, *7*(6), 591–605.

MacWhinney, B. (2014). *The Childes Project: Tools for Analyzing Talk, Volume I: Transcription Format and Programs*. Psychology Press.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* Cambridge, MA: MIT Press.

Margolis, E., & Laurence, S. (2012). In defense of nativism. *Philosophical Studies*.

Marotta, A., Lupiáñez, J., Martella, D., & Casagrande, M. (2012). Eye gaze versus arrows as spatial cues: two qualitatively different modes of attentional selection. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(2), 326–35.

Masangkay, Z. S., McCluskey, K. a, McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development*, *45*(2), 357–366.

Mayer, A., & Trauble, B. E. (2012). Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. *International Journal of Behavioral Development*, *37*(1), 21–28.

McAlister, A. R., & Peterson, C. C. (2013). Siblings, Theory of Mind, and Executive Functioning in Children Aged 3-6 Years: New Longitudinal Evidence. *Child Development*, *84*(4), 1442–1458.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(4474), 1139–1141.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631.

Michael, J., & Christensen, W. (2016). Flexible goal attribution in early mindreading. *Psychological Review*, *123*(2), 219.

Michael, J., & D'Ausilio, A. (2015). Domain-specific and domain-general processes in social perception – A complementary approach. *Consciousness and*

Cognition, *36*, 434–437.

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & Psychophysics*, *68*(2), 327–337.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–52.

Mill, J. S. (1865). *An Examination of Sir William Hamilton's Philosophy*. London: Longmans.

Miller, C. B. (2013). *Moral character: An empirical theory*. Oxford University Press.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–46.

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: the critical role of attitude diagnosticity of socially constrained behavior. *Journal of Personality and Social Psychology*.

Moll, H., Kane, S., & McGowan, L. (2015). Three-year-olds express suspense when an agent approaches a scene with a false belief. *Developmental Science*, n/a–n/a.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603–613.

Montgomery, D. E. (2005). The Developmental Origins of Meaning for Mental Terms. In *Why Language Matters for Theory of Mind* (pp. 106–122). Oxford University Press.

Moore, C., Bryant, D., & Furrow, D. (1989). Mental Terms and the Development of Certainty. *Child Development*, *60*(1), 167–171.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326.

Moriguchi, Y., Okanda, M., & Itakura, S. (2008). Young children's yes bias: How does it relate to verbal ability, inhibitory control, and theory of mind? *First Language*, *28*(4), 431–442.

Morton, A. (1996). Folk Psychology is not a Predictive. *Mind*, *105*(417), 119–137.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Clarendon Press/Oxford University Press.

Norenzayan, A., Choi, I., & Nisbett, R. (2003). Cultural similarities and differences in social inference: evidence from behavioral predictions and lay theories of behavior. *Human Resource Abstracts*, *38*(2).

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2010). Broca's Area and Language Processing: Evidence for the Cognitive Control Connection. *Language and Linguistics Compass*, *4*(10), 906–924.

Ogilvie, R., & Carruthers, P. (2016). The case against encapsulation. *Review of Philosophy and Psychology*, *7*.

Oh, S., & Lewis, C. (2008). Korean Preschoolers' Advanced Inhibitory Control and Its Relation to Other Executive Skills and Mental State Understanding. *Child Development*, *79*(1), 80–99.

Okanda, M., & Itakura, S. (2008). Children in Asian cultures say yes to yes--no questions: Common and cultural differences between Vietnamese and Japanese children. *International Journal of Behavioral Development*, *32*(2),

131–136.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–8.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, *107*(1), 179–217.

Papafragou, A., Cassidy, K., & Gleitman, L. (2007). When we think about thinking : The acquisition of belief verbs. *Cognition*, *105*(1), 125–165.

Pargetter, R. (1984). The scientific inference to other minds. *Australasian Journal of Philosophy*, *62*(2), 158–163.

Paulus, M., Hunnius, S., van Wijngaarden, C., Vrins, S., van Rooij, I., & Bekkering, H. (2011). The role of frequency information and teleological reasoning in infants' and adults' action prediction. *Developmental Psychology*, *47*(4), 976–983.

Pavan, G., Dalmaso, M., Galfano, G., & Castelli, L. (2011). Racial group membership is associated to gaze-mediated orienting in Italy. *PLoS ONE*, *6*(10).

Pearl, J. (2000). *Causality: models, reasoning, and inference*. New York, NY: Cambridge University Press.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–21.

Perner, J. (1991). *Understanding the representational mind. Learning, development, and conceptual change.* Cambridge, MA: MIT Press.

Perner, J. (2010). Who took the cog out of cognitive science? Mentalism in an era of

anti-cognitivism. In P. Frensch & R. Schwarzer (Eds.), *Perception, attention, and action: International Perspectives on Psychological Science (Volume 1)* (pp. 239–262).

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*, 125–137.

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of Mind Is Contagious: You Catch It from Your Sibs. *Child Development*, *65*(4), 1228–1238.

Perner, J., Sprung, M., Zauner, P., & Haider, H. (2003). Want That is Understood Well before Say That, Think That, and False Belief: A Test of de Villiers's Linguistic Determinism on German-Speaking Children. *Child Development*, *74*(1), 179–188.

Peterson, C. C., & Wellman, H. M. (2009). From fancy to reason: Scaling deaf and hearing children's understanding of theory of mind and pretence. *British Journal of Developmental Psychology*, *27*(2), 297–310.

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development*, *76*(2), 502–17.

Piantadosi, S. T., & Kidd, C. (2016). Extraordinary intelligence and the care of infants. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(25), 6874–9.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515.

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: evidence from learners of a new sign language. *Psychological Science*, *20*(7), 805–12.

Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*(03), 341–365.

Qureshi, A. W., Apperly, I., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: evidence from a dual-task study of adults. *Cognition*, *117*(2), 230–6.

Rakoczy, H. (2015). In defense of a developmental dogma: children acquire propositional attitude folk psychology around age 4. *Synthese*.

Rakoczy, H., Warneken, F., & Tomasello, M. (2007). "This way!", "No! That way!"—3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, *22*(1), 47–68.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Reeder, G. D. (2009). Mindreading: Judgments About Intentionality and Motives in Dispositional Inference. *Psychological Inquiry*, *20*(1), 1–18.

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional Attribution: Multiple Inferences About Motive-Related Traits. *Journal of*

Personality and Social Psychology, *86*(4), 530–544.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(34), 13526–31.

Ristic, J., Friesen, C. K., & Kingstone, A. (2002). Are eyes special? It depends on how you look at it. *Psychonomic Bulletin & Review*, *9*(3), 507–513.

Ristic, J., & Kingstone, A. (2005). Taking control of reflexive social attention. *Cognition*, *94*(3).

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–92.

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. *Advances in Experimental Social Psychology*.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science*, *24*(1), 27–33.

Rubio-Fernández, P., & Geurts, B. (2015). Don't Mention the Marble! The Role of Attentional Processes in False-Belief Tasks. *Review of Philosophy and Psychology*.

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review*, *34*(3), 265–293.

Ruffman, T., Perner, J., Naito, M., Parkin, L., & Clements, W. A. (1998). Older (but not younger) siblings facilitate false belief understanding. *Developmental*

Psychology, *34*(1), 161–174.

Ruffman, T., Slade, L., & Crowe, E. (2002). The Relation between Children's and Mothers? Mental State Language and Theory-of-Mind Understanding. *Child Development*, *73*(3), 734–751.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The Development of Executive Functioning and Theory of Mind. A Comparison of Chinese and U.S. Preschoolers. *Psychological Science*, *17*(1), 74–81.

Sabini, J., & Silver, M. (2005). Lack of Character? Situationism Critiqued. *Ethics*, *115*(3), 535–562.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in Black and White children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, *39*(4), 590–598.

Samson, D., Apperly, I., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255–1266.

Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., & Heyes, C. (2014). Avatars and arrows: implicit mentalizing or domain-general processing? *Journal of Experimental Psychology. Human Perception and Performance*, *40*(3), 929–37.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842.

Saxe, R., & Wexler, A. (2005). *Making sense of another mind: The role of the right*

temporo-parietal junction. *Neuropsychologia* (Vol. 43).

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433–438.

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, *23*(8), 842–7.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2014). What do we know about implicit false-belief tracking? *Psychonomic Bulletin & Review*.

Scholl, B. J., & Gao, T. (2013). Perceiving Animacy and Intentionality : Visual Processing or Higher-Level Judgment ? In M. D. Rutherford (Ed.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 197–230). MIT Press.

Scholl, B. J., & Leslie, A. M. (1999). Modularity , Development and " Theory of Mind ." *Mind & Language*, *14*(1), 131–153.

Scholl, B. J., & Leslie, A. M. (2001). Minds, modules, and meta-analysis. *Child Development*, *72*(3), 696–701.

Scott, R. M. (2014). Post hoc versus predictive accounts of children's theory of mind: A reply to Ruffman. *Developmental Review*, *34*(3), 300–304.

Scott, R. M., & Baillargeon, R. (2014). How fresh a look? A reply to Heyes. *Developmental Science*, *17*(5), 660–4.

Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: evidence from two novel verbal spontaneous-response tasks. *Developmental Science*, *15*(2), 181–193.

Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology*, *82*, 32–56.

Searle, J. (1975). Indirect Speech Acts. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (pp. 59–82). Academic Press.

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating Into the Future or Driven by the Past. *Perspectives on Psychological Science*, *8*(2), 119–141.

Senghas, A., Kita, S., & Ozyürek, A. (2004). Children creating core properties of language: evidence from an emerging sign language in Nicaragua. *Science*, *305*(5691), 1779–82.

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, *22*(7), 878–80.

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology*, *47*(5), 1239–1247.

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, *14*(3), 301–321.

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition*, *38*(1), 1–12.

Simons, M. (2007). Observations on embedding verbs, evidentiality, and

presupposition. *Lingua*, *117*(6), 1034–1056.

Slaughter, V., & Gopnik, A. (1996). Conceptual Coherence in the Child' s Theory of Mind: Training Children to Understand Belief. *Child Development*, *67*(6), 2967–2988.

Slessor, G., Laird, G., Phillips, L. H., Bull, R., & Filippou, D. (2010). Age-related differences in gaze following: does the age of the face matter? *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *65*(5), 536–41.

Smiley, P., & Huttenlocher, J. (1989). Young children's acquisition of emotion concepts. In C. Saami & P. L. Harris (Eds.), *Children's understanding of emotion* (pp. 27–49). Cambridge, UK: Cambridge University Press.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*(6), 907–12.

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, *130*(1), 1–10.

Spaulding, S. (2010). Embodied Cognition and Mindreading. *Mind & Language*, *25*(1), 119–140.

Spaulding, S. (2015). On Direct Social Perception. *Consciousness and Cognition*, *36*, 472–482.

Spaulding, S. (2016). Mind Misreading. *Philosophical Issues*, *26*.

Sperber, D., & Wilson, D. (2002). Pragmatics, Modularity and Mind-reading. *Mind and Language*, *17*(1&2), 3–23.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*(12), 1391–1408.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, *17*(3), 279–305.

Sreenivasan, G. (2002). Errors about Errors: Virtue Theory and Trait Attribution. *Mind*, *111*(441), 47–68.

Sripada, C. S. (2009). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, *151*(2), 159–176.

Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology*, *48*(1), 232–238.

Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language*, *26*(3), 353–380.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Stalnaker, R. (1998). On the Representation of Context. *Journal of Logic, Language and Information*, *7*(1), 3–19.

Steglich-Petersen, A., & Michael, J. (2015). Why Desire Reasoning is Developmentally Prior to Belief Reasoning. *Mind & Language*, *30*(5), 526–549.

Surian, L., & Leslie, A. M. (1999). Competence and performance in false belief understanding: A comparison of autistic and normal 3-year-old children. *British Journal of Developmental Psychology*, *17*(1), 141–155.

Surtees, A., Apperly, I., & Samson, D. (2013a). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, *129*(2), 426–438.

Surtees, A., Apperly, I., & Samson, D. (2013b). The use of embodied self-rotation for visual and spatial perspective-taking. *Frontiers in Human Neuroscience*, *7*(November), 698.

Surtees, A., Apperly, I., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition*, *150*, 43–52.

Surtees, A., Butterfill, S., & Apperly, I. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *The British Journal of Developmental Psychology*, *30*(Pt 1), 75–86.

Surtees, A., Samson, D., & Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97–105.

Symons, D. K. (2004). Mental state discourse, theory of mind, and the internalization of self–other understanding. *Developmental Review*, *24*(2), 159–188.

Symons, D. K., Fossum, K.-L. M., & Collins, T. B. K. (2006). A Longitudinal Study of Belief and Desire State Discourse During Mother?Child Play and Later False Belief Understanding. *Social Development*, *15*(4), 676–692.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). *Moral signals, public outrage, and immaterial harms*. *Journal of Experimental Social Psychology* (Vol. 47).

Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology*, *36*(1), 25–43.

Taumoepeau, M., & Ruffman, T. (2006). Mother and Infant Talk About Mental States

Relates to Desire Language and Emotion Understanding, *77*(2), 465–481.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., Goodman, N. D., Xu, F., Tenenbaum, J. B., … Lengyel, M. (2011). How to grow a mind: statistics, structure, and abstraction. *Science (New York, N.Y.)*, *331*(6022), 1279–85.

Terrizzi, B. F., & Beier, J. S. (2016). Automatic cueing of covert spatial attention by a novel agent in preschoolers and adults. *Cognitive Development*, *40*, 111–119.

Teufel, C., Alexis, D. M., Clayton, N. S., & Davis, G. (2010). Mental-state attribution drives rapid, reflexive gaze following. *Attention, Perception & Psychophysics*, *72*(3), 695–705.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J.-R. (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), 4567–70.

Thompson, J. R. (2014). Signature Limits in Mindreading Systems. *Cognitive Science*, *38*(7), 1432–1455.

Thomson, J. J. (1976). Killing, Letting Die, and the Trolley Problem. *Monist*, *59*(2), 204–217.

Todorov, A. (2013). Making up your mind after 100-ms exposure to face. *Psychological Science*, *17*(7), 592–598.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard University

Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *The Behavioral and Brain Sciences*, *28*(5), 675–91; discussion 691–735.

Tomasello, M., & Rakoczy, H. (2003). What Makes Human Cognition Unique? From Individual to Shared to Collective Intentionality. *Mind and Language*, *18*(2), 121–147.

Trope, Y., & Gaunt, R. (2007). Attribution and person perception. In M. Hogg & J. Cooper (Eds.), *The Sage handbook of social psychology* (pp. 176–194). London: Sage Publications.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I Know What You Are Doing. *Neuron*, *31*(1), 155–165.

Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child Development*, *81*(6), 1661–9.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858.

Warneken, F., & Tomasello, M. (2007). Helping and Cooperation at 14 Months of Age. *Infancy*, *11*(3), 271–294.

Warneken, F., & Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences*, *13*(9), 397–402.

Wellman, H. M. (2012). Theory of mind: Better methods, clearer findings, more development. *European Journal of Developmental Psychology*, *9*(3), 313–330.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford: Oxford University Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, *72*(3), 655–84.

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science*, *17*(12), 1075–1081.

Wellman, H. M., Fuxi, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: longitudinal perspectives. *Child Development*, *82*(3), 780–92.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541.

Wellman, H. M., & Peterson, C. C. (2013). Deafness, thought bubbles, and theory-of-mind development. *Developmental Psychology*, *49*(12), 2357–67.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245–275.

Westra, E. (2016a). Pragmatic development and the false-belief task. *Review of Philosophy and Psychology*, 1–23.

Westra, E. (2016b). Spontaneous mindreading: a problem for the two-systems account. *Synthese*.

Westra, E. (2017). Character and theory of mind: an integrative approach.

Philosophical Studies, 1–25.

Westra, E., & Carruthers, P. (2017). Pragmatic development explains the Theory-of-Mind Scale. *Cognition*, *158*, 165–176.

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: how attentional selection is shaped by ascribing intentions to others. *PloS One*, *7*(9), e45391.

Wilson, D., & Sperber, D. (2012). *Meaning and Relevance*. Cambridge University Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, *5*, 674.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5012–5.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(20), 8235–

40.

Zawidzki, T. W. (2011). How to Interpret Infant Socio-Cognitive Competence. *Review of Philosophy and Psychology*, *2*(3), 483–497.

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, *193*(12), 3951–3985.