FROM TWO SYSTEMS TO A MULTI-SYSTEMS ARCHITECTURE FOR MINDREADING[1]

Wayne Christensen (corresponding author)
Macquarie University, University Avenue, Macquarie Park, Sydney 2113, Australia
Email: wayne.christensen@gmail.com
Phone: 00 61 498 109157

&

John Michael
Department of Cognitive Science, Central European University, Budapest, Hungary
Email: johnmichaelaarhus@gmail.com

[1] Thanks to the special editor Marcin Miłkowski, and to Ágnes Kovács and Dora Kampis for helpful comments. John Michael's work was supported by a Marie Curie Intra-European Fellowship (grant nr: 331140) within the framework FP7-PEOPLE-2012-IEF.

Abstract

This paper critically examines Apperly and Butterfill's parallel 'two systems' theory of mindreading and argues instead for a cooperative multi-systems architecture. The minimal mindreading system (system 1) described by Butterfill and Apperly is unable to explain the flexibility of infant belief representation or fast and efficient mindreading in adults, and there are strong reasons for thinking that infant belief representation depends on executive cognition and general semantic memory. We propose that schemas, causal representation and mental models help to explain the representational flexibility of infant mindreading and give an alternative interpretation of evidence that has been taken to show automatic, fast and efficient belief representation in adults.

Highlights

- We argue that Butterfill and Apperly's minimal mindreading account fails to explain the flexibility of infant mindreading.
- A cooperative multi-system architecture can better explain the flexibility of infant mindreading than a parallel two systems architecture.
- Schemas, causal representation and mental models can help to explain the representational flexibility of efficient mindreading.
- We predict that fast and efficient mindreading processes will be susceptible to top-down influences.

Keywords

social cognition, theory of mind, development, false belief, representation

1 Introduction

In a standard false belief task, a child (or other test subject) observes as an agent places an object (sometimes a doll) in a box at location A and then temporarily departs, whereupon some other agent arrives on the scene and transfers the object to a box at location B. When the first agent returns, the child is asked where the first agent (Sally) is likely to search for the object.[2] The correct answer, of course, is that Sally is likely to search in the box at location A, because that is where she falsely believes the doll to be located (Wimmer and Perner, 1983; Baron-Cohen et al., 1985; Call and Tomasello, 2008). This task has been regarded as a litmus test for the capacity to represent the beliefs of other agents because the child can't use her own knowledge of the location of the doll to predict where Sally will search; the child must distinguish Sally's belief about the location of the doll from the actual location. It has been a robust finding that children under the age of four years tend to fail at the standard false belief task, which prompted a widespread view that children younger than about four don't represent beliefs in others. This view has come under pressure, however, from recent evidence that children can succeed on various modified versions of the task that don't require a verbal response by around their first birthdays (e.g. Onishi and Baillargeon, 2005; Surian et al., 2007; Southgate et al., 2007; Song, Onishi, & Baillargeon, 2008; Buttelmann et al., 2009; Baillargeon et al., 2010; Träuble et al., 2010), or indeed even by seven months (Kovacs et al., 2010).

Thus, with evidence of mindreading in early childhood now having been reported by researchers from various labs using diverse methods, the theoretical challenge is to reconcile the large discrepancy in results between verbal false belief tasks and non-verbal versions. As we shall see, this puzzle raises many complex, difficult questions not only about the development of mindreading but also about the nature of mental representation and cognitive architecture. In this paper we'll examine in some detail one attempt to resolve the puzzle: Apperly and Butterfill's 'two systems' account of mindreading (Apperly and Butterfill, 2009; Apperly, 2011; Butterfill and Apperly, 2013). This account addresses many of these issues in insightful ways, providing a

[2] The experiments are easier to interpret when described concretely, and in what follows we will refer to the first and second agents as Sally and Anne, and the object as a doll, except where there are variations and the specific details matter.

characterization of key representational differences between infant and adult mindreading that is linked to an analysis of a tension between efficiency and flexibility, which in turn provides an explanation for why there should be two mindreading systems.

Our discussion identifies a number of points of agreement with Apperly and Butterfill. We think they are right that multiple systems contribute to mindreading, and that some of these systems employ representations that lack the full structure of adult psychological concepts. We also agree that the nature of these systems and their relations is fundamentally shaped by a trade-off between efficiency and flexibility. But we argue for a different architectural solution to these constraints. Apperly and Butterfill believe that the competing demands of efficiency and flexibility give rise to a *parallel two systems architecture* for mindreading. Specifically, they believe that the tension between efficiency and flexibility is such that the only way that both can be achieved in mindreading is by means of distinct cognitive systems. Moreover, because efficiency is attained by means of hard constraints, the efficient system should be capable of only limited information exchange with the flexible system, and the two systems should consequently operate largely in parallel. We'll argue that this is oversimplified, and propose instead a *cooperative multi-system architecture* for mindreading. Efficiency can be compatible with rich information exchange amongst multiple cognitive systems, and a cooperative multi-system organisation can yield a better overall balance of efficiency and power than a parallel architecture. Furthermore, recent evidence for infant mindreading suggests that it involves a kind of flexibility that is better explained by a cooperative multi-system architecture.

We also criticize the assumptions that Apperly and Butterfill make concerning the representational characteristics of their two systems. On the one hand, they propose that efficient mindreading is achieved by a simple, inflexible representational scheme, and on the other hand, they propose stringent requirements on the belief representation performed by the flexible mindreading system, which they see as representing beliefs 'as such'. We argue that the motivations for these proposals are misguided. Specialized cognitive systems can be capable of relatively complex forms of representation, and flexibility can be achieved through cooperative multi-system interactions. We also argue that flexible conceptual belief representation can be

simpler and more heterogeneous than Apperly and Butterfill suppose. Infants probably do not represent beliefs 'as such' in the way that older children and adults do, but it's likely that their belief representations involve generalized semantic memory, and that they are developing forms of conceptual belief representation that serve as a basis for the more sophisticated forms of belief representation that emerge in older children. This learning process probably involves the acquisition of a rich stock of schemas and a constructivist progression in which schemas are refined and more abstract conceptualizations are developed.

2 Evidence for mindreading in infants

Evidence for the representation of false belief in infants first arose from experiments that used looking behavior instead of verbal response as the measure indicating the presence of the representation. Clements and Perner (1994) found an intriguing inconsistency in young children's responses in a false belief situation: 90% of the children between 35 months and four-and-a-half years old looked first to the empty location (where the Sally falsely believed the object to be located), and yet only 45% of them gave the correct verbal answer when asked where Sally was likely to search. The authors speculated that the children's anticipatory looking might indicate that they implicitly represented Sally's false belief.

This result foreshadowed a new approach to the false belief paradigm focusing on non-verbal measures, especially looking behaviour.[3] Onishi and Baillargeon (2005) found that children looked longer when Sally searched in the object's actual location compared with when Sally searched in the original (incorrect) location. According to the authors, the increased looking time indicated a violation of expectation, which revealed that the child had implicitly formed the expectation that Sally would search in the wrong location because she had a false belief. This finding has since been corroborated by numerous groups using similar paradigms: Surian and colleagues (2007), for example, observed the same pattern in a study involving 13 month-olds; Kovacs et al. (2010) found similar evidence in a study with seven-month-olds; and Southgate et al (2007), measuring children's anticipatory looking as in the Clements

[3] It is common to refer to experimental measures that depend on verbal responses as 'explicit' and measures based on behavior as 'implicit'. We will avoid this, however, because we think that it shouldn't be inferred that cognitive processes are implicit if they manifest in behavioral but not verbal measures. See also Carruthers (2013, p. 145), who makes a similar point.

and Perner (2004) study, found evidence that 25 month-old children first looked toward the wrong location upon Sally's return (after Sally had briefly looked away and failed to witness that a ball was transferred from one box to another), apparently in anticipation that Sally would search for the ball at the wrong location.

Whether these experiments reveal belief representation has been controversial.[4] But converging results have been reported in recent years using a variety experimental paradigms and measures. Thus, some experiments have probed whether infants are sensitive to variations in the epistemic conditions that produce belief. Senju et al. (2011) gave 18-month-olds experience with either an opaque or a transparent "trick" blindfold, then showed them a version of the false-belief scenario in which Sally wore the blindfold that the child had just experienced, and so either could (with the trick blindfold) or could not (with the opaque blindfold) see the transfer of the doll. The children showed anticipatory looking consistent with false belief representation only for the opaque blindfold.

Other studies have found belief sensitivity in infants and young children based on non-visual sources of information. In a study involving 15-month-olds, Träuble et al. (2010) used an apparatus consisting of a balance beam with a box at each end. When a foam ball was placed in one of the boxes, and that end of the beam was raised, the ball would noiselessly roll to the other box. Three conditions were compared. In a *true belief* condition Sally manipulated the beam herself while facing forward and able to see the transfer of the ball. In a *false belief* condition Sally's back was turned and the beam was manipulated without her input, resulting in the transfer of the ball. In a *manipulation* condition Sally again faced away from the apparatus, but she manipulated the beam herself, causing the ball to transfer between boxes. In each condition, two different outcomes were contrasted: either Sally reached for the ball in the original box (wrong location) or in the new box into which the ball had rolled (correct location). Träuble et al. found that in the false belief condition the infants looked longer when Sally reached for the correct location, whereas in the other two conditions the pattern was reversed, with the infants looking longer when Sally reached for the wrong location. Thus, the infants appeared to expect Sally to have a

[4] For discussions see Haith and Benson (1998), Kagan (2008), Müller and Giesbrecht (2008), Heyes (2014), and Ruffman (2014).

true belief in the manipulation condition even though she had not seen the object transfer because her back had been turned. This indicates that infant belief sensitivity can take into account varied informational sources for beliefs, including visual perception and non-visual object manipulation.

Similarly, Song, Onishi & Baillargeon (2008) found that 18-month-old infants' expectations in a false belief task were influenced by communication between the experimenter and Sally. In one condition the experimenter told Sally that the doll had been moved, and the infants did not then expect Sally to approach the original box. In another condition the experimenter merely told Sally that she liked the doll, but did not tell Sally that it had been moved. In this case the infants did expect Sally to have a false belief about the location of the doll. This indicates that the infants were able to take into account the specific content of the communication in adjusting their expectations for Sally's actions, and not just that there was communication concerning the ball.

In addition to false belief about location, there is also evidence that infants in the second year of life are able to attribute other reality-incongruent informational states to agents (for a review, see Baillargeon et al., 2010). For example, Scott et al. (2010) reported evidence that 18 month-olds could attribute a false belief to Sally concerning non-obvious properties of an object, namely which of two test objects could be shaken to produce a rattling sound.

Measures other than anticipatory looking have also indicated belief sensitivity in infants and young children. Buttelmann and colleagues (2009) used an 'active helping' paradigm in which 18 month-olds had to take into account Sally's false belief in order to determine her goal and help her to attain it. The paradigm employs the standard false belief scenario in which Sally returns to the scene, unaware that during her absence the doll has been moved from one box to another. Sally then tries but fails to open the box that doll was in originally (but is not now). The key finding was that most children did not attempt to help Sally open the box she was struggling with. Instead, they went over to other box and retrieved the doll. In contrast, in the true belief condition, where Sally had witnessed the transfer of the doll, most children assisted her with the box she was trying to open.

A very different measure was used in a study involving 17 month-olds, conducted by Southgate, Chevallier, & Csibra (2010). In this experiment Sally placed two different toys in two boxes and then departed. During her absence Anne switched the two toys. When Sally returned, she pointed to one of the boxes and said that the toy inside that box was a 'sefo'. When the infants were subsequently prompted to get the sefo, most of them retrieved the toy in the other box (the one Sally had not pointed to). This strongly suggests that the infants had taken Sally to hold a false believe about the locations of the toys, and intended to refer to the toy in the other box as a sefo.[5] More recently, Southgate and Vernetti (2014) conducted an EEG study in which they recorded sensorimotor alpha suppression in 6-month-olds when observing Sally confronted with a box in which she falsely believed a ball to be located. No such alpha suppression was recorded in a condition in which Sally correctly believed the box to be empty. The authors argue that this pattern of findings constitutes evidence that infants generate differential predictions about whether Sally will perform an action depending on that agent's beliefs about objects' locations.

3 Problems with deflationary and nativist explanations

Taken as a whole, these findings reveal an impressive specificity and flexibility in infant belief sensitivity which suggests that infants represent beliefs in some way. To provide a context for understanding Apperly and Butterfill's proposal for how infants represent beliefs we'll first briefly characterize and argue against two alternative approaches. Apperly and Butterfill's theory constitutes one kind of intermediate position between two extremes: deflationary accounts which claim that infant abilities can be explained without appeal to the representation of belief, and nativist accounts which propose that infants have an innate, fully-formed belief concept. Each of these extremes has problems, and the nature of these problems provides motivation for exploring the middle ground that Apperly and Butterfill have identified. Our criticisms of their account in later sections will show that another type of intermediate position is possible and has advantages, but we first need to give reasons for moving to the middle ground.

[5] Carpenter, Call, and Tomasello (2002) had earlier reported a very similar finding with 2-3 year-old children, and interpreted it as evidence that children take pragmatic contexts and others' mental states into account in learning new words.

3.1 Problems with deflationary accounts

The difficulties confronting deflationary approaches to infant successes on belief tasks are exemplified by problems with the 'behavior rule' account (Perner and Ruffman, 2005; Ruffman 2014). On this view, the abilities shown by infants in false belief tests can be explained as the result of behavior reading based on statistical learning, rather than the representation of beliefs. Thus, whereas a mindreading account claims that infants represent *behavior → mental state → behavior* relations, a behavior reading account proposes that infants represent *behavior → behavior* relations. This has the prima facie virtue of being in keeping with Morgan's Canon, the principle that psychological explanations should favor the simplest cognitive mechanism consistent with the evidence (Morgan, 1903). But the approach has been criticized on the grounds that it is unparsimonious when elaborated to account for the evidence that infants can predict belief-dependent actions in varied conditions. The behavior reading view must postulate a very large number of seemingly ad hoc behavior rules, whereas a mindreading account can more economically explain the ability of infants to predict belief-dependent actions in varied conditions by postulating that they represent beliefs (e.g., Song et al., 2008; Carey 2009; Apperly, 2011).

This criticism identifies a crucial issue in assessing the respective merits of deflationary and nativist accounts of apparent belief representation in infants, but whether the problem is decisive depends on the details. Ruffman (2014) argues that we have independent reasons to believe that statistical learning plays a powerful role in infant cognitive development, and if we allow that behavior rules can incorporate generalization, then a relatively small and plausible number of rules can account for infant abilities. But although it is undoubtedly pervasive and important, statistical learning has limitations which make it unlikely that it is the main basis upon which infants interpret intentional action. Causal learning should be distinguished from statistical learning, and the latter has key advantages that have implications for the representation of intentional action by infants.

By *causal representation* we mean representation that attributes causal powers to entities in the world, and represents relations and processes governed by those causal

powers. On this definition causal representation is mechanistic.[6] In contrast, by *statistical representation* we mean representation that captures correlations but does not represent the mechanistic basis for the correlations. Thus, a statistical representation of an $X \rightarrow Y$ relation only expresses a correlation between X and Y, while a causal representation of an $X \rightarrow Y$ relation identifies a mechanistic basis for the correlation.

For example, here is a causal description of a simple scenario. A small ball rolls down a sloping surface and passes through a narrow gap in a barrier, followed by a large ball that rolls down the slope and is blocked at the barrier because it is too big to pass through the gap. A purely statistical description would not mention balls, the barrier, physical movement, or blockage, because these are causal entities, processes and events. It would simply depict a temporally extended complex perceptual pattern. A hybrid representation that includes entities located in space, but not causal processes, would individuate the slope, the balls and the barrier, but include no mechanistic information concerning the structure of the events. The balls do not roll down the slope (this is causal); they simply exhibit spatial displacement. The small ball does not pass through the narrow gap because it is smaller than the gap; there is simply a particular spatial transition. The large ball is not blocked by the barrier because it is too big to fit through the gap; it simply manifests a different pattern of spatial movements to the small ball.

This example helps to reveal some of the advantages of causal representation in comparison with statistical representation. These include providing a structured basis for *salience*, *generative model construction*, and *diagnosis*. Causal representation provides a structured basis for salience by specifying the key causal properties present in a situation. Thus, the spherical shape of the balls, the size of the balls, the angle of the slope, the barrier, and the size of the gap in the barrier are all important causal properties present in the scenario just described. An observer who attends to the causal properties present in the situation can use them as the basis for constructing a *model* that *interprets* the situation (Johnson-Laird 1983). An interpretive model

[6] Importantly, causal representation as we are defining it is not just representation constrained in such a way as to respect causal relations, as is the case, e.g., with a Bayesian causal model (Pearl 2000). Rather, causal properties such as shape, size and hardness are represented explicitly as attributes of entities.

specifies the key relations present in the situation that govern what happens. An interpretive model can be used post hoc to understand a situation that has occurred, but it can also be used for predictive, hypothetical, and counterfactual judgments. In the case of prediction, the model allows the observer to anticipate the outcome before it happens. In the case of hypothetical judgment, the model allows the individual to determine what will happen if particular conditions obtain. An especially important form of hypothetical judgment is determining what action or actions should be taken to achieve a particular goal. Counterfactual judgments allow the individual to determine what might have occurred if the conditions were different.

Causal representation provides a basis for generative model construction because known entities and causal properties can be combined in novel ways to yield models for situations that haven't been previously experienced. There are of course limitations to generative model construction, and an observer attempting to interpret a situation that is relatively unfamiliar may be unable to construct a successful model, possibly because the observer isn't aware of all of the important causal features of the situation, or because she is unable to identify the important relationships. Maier's (1931) 'pendulum problem' illustrates the difficulties that can occur in constructing an integrated model for an unfamiliar situation that allows the individual to achieve a goal. Participants were introduced to a room that contained a variety of objects and two strings hanging from the ceiling. The task was to tie the strings together, but they were spaced such that it was impossible to hold one string and reach the other. A solution is to tie an object to one of the strings and swing it as a pendulum, so that the movement brings it within reach when the other string is held. However, few participants discovered this solution.

Familiar situations are easier to interpret than novel situations because learning organizes knowledge in a way that tends to make key properties and relations more apparent. The participants in Maier's study – adults – presumably had all the background causal knowledge required for solving the task, but they were unable to put the information together in a way that revealed the solution. One of the primary explanations for the ability to organize related information is by means of schemas, which are abstract structures that capture stereotypical relations (Bartlett, 1932; Piaget, 1952; Schank and Abelson, 1977; Gureckis et al. 2010). Thus, a young child

will have schemas for commonly experienced situations like mealtimes, playing with toys, greetings and saying goodbye, birthday parties, bedtime, and so on. Schemas can facilitate the construction of a model of a situation by providing an organizing framework that integrates information and highlights key relationships (Kintsch and van Dijk 1978), and thereby also facilitates predictions about how the situation will develop (or would develop in counterfactual cases).

In cases where an agent lacks an adequate schema to pick out and organize the relevant causal features in a situation, she will accordingly fail to predict how the situation develops. But causal knowledge required in order to interpret the situation may nevertheless be available – albeit not yet organized into an adequate schema. This causal knowledge can be harnessed for post hoc *diagnosis* of the situation. An infant observing the ball scenario described above might fail to predict the blocking of the large ball because she fails to construct an integrated causal model of the situation that captures the relevant causal attributes. The stopping of the large ball at the barrier is a surprising event for the infant, but she can use causal knowledge to diagnose the cause of the event: the large ball is too big to fit through the gap. Now attuned to the relevant causal parameters, if she is presented with the scenario again she can rapidly construct a causal model that captures the key relations, and predict the event before it happens. Indeed, when attuned to the appropriate causal parameters the infant should be able to successfully predict outcomes in variations of the scenario in which the sizes of the balls and the barrier are altered.

Pure statistical representation doesn't provide this kind of causally focused diagnosis and generative prediction. But, on the other hand, because statistical learning doesn't depend on causal presuppositions, correlations can be learned even when no causal model of the relation can be readily constructed. Our causal understanding of the world develops slowly and is highly incomplete (even for very knowledgeable adults), so statistical learning plays an essential role in the discovery of patterns. Based on these considerations, then, we should expect that statistical and causal learning both play important roles in infant cognitive development, and that they are mutually supportive. The detection of statistical patterns can guide causal learning, and conversely, the causal individuation of important properties, entities, relations and

processes provides structure on which statistical learning can operate. Indeed, the formation of schemas will involve both causal and statistical learning.

This line of reasoning has implications for the parsimony argument against the behavior rules approach. Given the power and usefulness of causal representation for constructing interpretive models, it is unlikely that infants primarily or exclusively relay on statistical behavior rules to interpret the actions of other agents. Ruffman (2014, pp. 275-6) argues that a 'mentalist' account of infant belief sensitivity is actually less parsimonious than a behavior reading account because the mentalist account must take into account all of the complexity of the conditions that influence the behavior the infant is predicting (in the standard task, where Sally will search), just as a behavior rules account must, and in addition proposes that the infant represents Sally's beliefs. But Ruffman fails to appreciate the distinctive predictive power and economy of model-based representation in comparison with statistical rules. Firstly, models provide a flexible basis for integrating complex situational information. Thus, novel configurations and items of information that are acquired over time can be meaningfully connected. Secondly, models provide an economical basis for prediction. Every variation in the situational factors that results in a significantly different outcome requires a distinct statistical rule, whereas causal representation can more economically cope with situational variation because it permits the flexible construction of models that capture the structure of a given situation and yield predictions for that situation. A rule appropriate for the situation isn't needed in advance because a model can be constructed 'on the fly'.

The ability to cope with situational variation generatively is a major advantage for model-based representation because any system that relies entirely on pre-specified rules will face a problem of combinatorial explosion: every variation in the situation that makes a significant difference to the outcome requires a distinct rule, but as the number of factors in the situation increases linearly the number potential situational configurations increases exponentially.

Model-based representation of the actions of other agents need not incorporate the representation of belief, and indeed causal models can serve as the basis for generating statistical predictions. However, there are reasons for thinking that infants

will use model-based representation to differentiate key aspects of the intentional structure of action and agents. This is because the causal structure of action is very sensitive to its intentional features. Changes in factors like motivation, goals, attention, physical capacities, and environmental constraints and opportunities profoundly affect which actions an agent will perform and how she will perform them. There is now a substantial body of evidence that infants are sensitive to these kinds of factors and can flexibly integrate diverse kinds of agency information (Woodward, 2005). For example, Csibra et al. (2003) employed an agentic scenario similar to the non-agentic causal scenario described above. 6 and 12-month-old infants were shown a scene in which a small ball appears to be chased by a large ball. The small ball passes through a narrow gap in the barrier, while the large ball moves around the barrier. 6 and 12-month-old infants expect the large ball to catch the small ball and are surprised by a final scene in which they are separate. In a variation, the gap in barrier is wide enough for the large ball to pass through, and 12-month-old infants are surprised if the large ball nevertheless goes around the barrier. This suggests that 12-month infants can take both goals and the structure of the environment into account in interpreting action.

If we allow that infants are differentiating key features of intentional action, and flexibly integrating agentic information in situation-specific ways, this lends presumptive weight to the view that infants succeed on false belief tasks because they represent beliefs. Note that evidence for successful prediction that tracks variations of a key parameter in novel situations is strongly suggestive that the parameter is represented and integrated into a situation model. Ruffman claims that Senju et al.'s (2011) 'trick blindfold' experiments, described in section 2, can be explained on the basis that infants recognize what an agent does and does not see, and employ the behavior rule that an agent will search for an object at an original location only when she hasn't seen it subsequently moved. The experience of the infants with the trick and ordinary blindfolds allows them to appropriately adjust their expectations for what the agent can see. This is possible, but even on this interpretation infants have a rich understanding of 'seeing' that includes a causal understanding that seeing provides information for the agent.

Moreover, this rule can't explain the ability of infants to predict Sally's behavior when she has been told where the ball is (Song et al., 2008). Ruffman claims that this result can be explained on the assumption that infants know that verbal communication can indicate location for themselves, and they expect other agents also to use communication to indicate location. But, crucially, this understanding of 'indication of location' is not linked to any particular behavior, so it is hard to see how it differs from the representation of belief. If infants understand that another agent can use highly varied sources of information to 'indicate location', that this information can be retained, and that the agent can use the information (which may be incorrect) to guide varied behavior, this essentially amounts to representing beliefs of the agent concerning location.

3.2 Problems with nativist approaches to infant belief competencies

The problems of the deflationary approach lend weight to the view that infants really are representing beliefs. This raises the possibility that a nativist interpretation of the evidence may be correct. In other words, the early emergence of belief sensitivity – some studies have found belief sensitivity at 6-7 months (Kovacs et al., 2010; Southgate and Vernetti, 2014) – may be taken to reveal that belief representation is based on an innate belief concept (e.g. Baillargeon et al., 2010; Leslie, 2005; Carruthers, 2013). The central problem for nativism lies in explaining why children struggle specifically with explicit verbal false belief tasks for years after succeeding in paradigms using implicit measures. One possible explanation is that young children struggle to overcome a default assumption that beliefs are true (Leslie et al. 2004). Thus, when predicting an agent's behavior, they have a tendency to ascribe to the agent whatever relevant knowledge they themselves have, and in so doing succumb to the 'the curse of knowledge' (Birch and Bloom, 2007; Carlson and Moses, 2001). On this view, children must first develop a capacity for executive control in order to inhibit their own perspective and thereby overcome the 'curse of knowledge'.

There are, however, compelling reasons to doubt that this could be the whole story. First, Call and Tomasello (1999) reported that children performed no better in an experiment in which the potential reality bias was removed, i.e. children needed to take into account an agent's false belief, but they themselves did not know the true location of an object. Secondly, there is evidence that Chinese children outperform

American children in executive function but not in mindreading (Sabbagh et al. 2006). Thirdly, Yott and Poulin-Dubois (2012) found a strong correlation between looking time in the test condition of a non-verbal false belief task and success on an executive function task in which the 18 month-olds had to inhibit a prepotent response (reaching for a toy through a transparent window) in order to carry out a desired response (turning a dial on the side of the box). As the authors emphasize, this is difficult to reconcile with the conjecture that mindreading in infants is performed by a system that operates on an automatic basis, i.e. which does not require executive control. Fourthly, a recent longitudinal study conducted by Thoermer et al. (2012) found evidence that casts doubt upon nativist interpretations. Specifically, they found that 18 month-olds who passed an implicit false belief task were more likely to pass an explicit verbal false belief task about locations at four but were not more likely to pass an explicit verbal false belief task about contents. Thus, the longitudinal continuity was *task-specific.* This is consistent with the possibility that children gradually acquire familiarity with specific situations that they frequently encounter, and that their performance improves most quickly in situations that they frequently encounter and less quickly in situations that they encounter less frequently. But this is not easy to reconcile with nativist interpretations of the infant mindreading data: according to nativism, the capacity that is tapped in implicit mindreading tasks in the 2nd year (or earlier) is the capacity to represent beliefs in general (i.e. it is the belief concept), so it is not clear why the continuity between the early-emerging capacity and later capacities should be specific to particular tasks

More generally, belief representation in older children and adults involves rich explicit semantic knowledge about perception, memory, beliefs, reasoning, etc. There is evidence that language experience contributes to an understanding of perspective differences (Harris 2005), and an understanding of mental states and communication in context (Dunn and Brophy 2005), and that there are coordinated improvements in the understanding of beliefs and other mental states, and representation more broadly, around ages 3-4 (Apperly, 2011). This suggests that the delayed ability to succeed on the standard explicit false belief task depends on qualitative changes in representational ability, not just improvements in mindreading efficiency.

4 Apperly and Butterfill: two systems for mindreading

There are thus good reasons for taking an approach to the development of mindreading which postulates that multiple systems and forms of representation are involved. Apperly and Butterfill's two system account is consequently of particular interest because it occupies a promising region of the theoretical space, and it illuminates a number of pressing issues. They propose that the disparity between evidence for early mindreading and delayed success on the standard false belief task is explained by the fact that there are two systems for mindreading: an early developing system that employs simple representations to achieve fast, resource-efficient processing, and a later-developing system that depends on language and executive resources to achieve powerful, flexible processing (Apperly and Butterfill, 2009; Apperly, 2011; Butterfill and Apperly 2013).

In support of this idea they appeal to number cognition as a parallel. In this case too there is a contrast between simple competencies found in infants (and nonhuman animals) and more complex abilities that develop gradually in older children. Infants as young as 5 months old are sensitive to the number of items in arrays of objects, but their numerical abilities show strong limitations: there is a capacity limit of 3-4 items for tracking precise numerosity, and they need a large ratio of difference in order to judge size differences between collections with large numbers of items. The marked differences between the early and later developing abilities support the idea that there are distinct systems involved in number cognition: a relatively simple system which develops early, and a more powerful system that develops later and is dependent on training in mathematical concepts such as the number system (see Carey, 2009, chapter 4, for an overview of this research).

The overall similarity with the developmental pattern of mindreading abilities suggests that there may likewise be two substantially distinct mindreading systems. To give theoretical support to this idea, Apperly and Butterfill argue that there is a fundamental tension between flexibility and efficiency, with simple representational capacities providing efficiency at the expense of flexibility, and powerful representational capacities providing flexibility at the expense of efficiency. Infant mindreading occurs in a context of very limited general cognitive resources, and cannot draw on the representational capacities that linguistic experience makes

available for later mindreading. Conversely, Apperly and Butterfill point out that adult-like mindreading is very flexible, involving in some cases complex inferences that are likely to depend on substantial cognitive resources. If efficiency and flexibility cannot be reconciled in one system, then the two kinds of mindreading must be supported by at least two distinct cognitive systems.

Butterfill and Apperly (2013) give a detailed characterisation of the representational characteristics of the efficient system, which they characterise as a system for 'minimal mindreading'. Their intention is to specify a representational system which can track belief in simple situations, but which does not incorporate psychological concepts. By design this system does not represent beliefs 'as such', and it lacks the flexibility that Butterfill and Apperly associate with adult mindreading. The system includes four principles for interpreting the actions of other agents, which are as follows.

(1) The first principle is that bodily movements form units that are directed to towards goals (p. 614). Butterfill and Apperly define goals as the teleological outcomes of bodily movements, where this means that the movements are organized in order to bring about the outcome (p. 613). This principle allows the individual to infer goals from action without representing them in terms of psychological states involving intentions, beliefs, and other propositional attitudes (p. 613).

(2) The second principle employs two representations – *field* and *encountering* – that together serve as a simplified surrogate for perception. Butterfill and Apperly say that a field is a "set of objects" (p. 614), but they appear to mean that it is an area specified in relation to the agent that encompasses objects. They say that the agent's field is determined by factors such as proximity, lighting, eye direction, barriers, and so on. Encountering is a relation between the agent and an object, and it occurs when the object is in the agent's field. The second principle is that the agent must encounter an object before she can engage in goal-directed actions aimed at the object (p. 615).

(3) This principle employs a concept of *registration* that serves as a partial surrogate for belief. Having encountered an object in a particular location, an agent registers it as being in that location, and will continue to register it as in that location until she

encounters it in a different location (p. 617). Registration consequently "has a correctness condition which may not obtain" (p. 617), because the object may have moved since the agent's last encountering. The third principle in full is that an agent must correctly register an object as in a particular location if she is to successfully perform a goal directed action aimed at the object. Butterfill and Apperly say that two kinds of inference can be drawn using this principle. An agent who has not correctly registered an object will not be able to successfully perform actions aimed at the object. And if an agent does succeed in performing an action with the object as a goal, she has correctly registered its location (p. 617).

(4) The fourth principle is that an agent will direct actions that have the object as a goal to the location at which she has registered the object (p. 619). With this principle an infant can predict that Sally will search in the empty box (p. 620).

Butterfill and Apperly claim that these principles can allow an individual to track beliefs in a limited but useful range of circumstances without requiring that they be represented 'as such'. Apperly and Butterfill (2009) characterize the latter form of belief representation in terms of the representation of beliefs as propositional attitudes. Specifically, they say that the representation of belief 'as such' involves representing it as an attitude to "a content" that plays a certain psychological role. They describe the content as 'propositional', which they define as 'sentence-like' (2009, p. 957), and as allowing for beliefs with complex contents, such as those involving quantification (2009, p. 960). They describe the psychological role of belief as including being caused and justified by perception, as interacting with other beliefs and desires, and as causing and justifying actions (2009, p. 957). They also claim that adult-like belief attribution involves abductive inference, and can draw on an unlimited range of information (2009, p. 960). Based on the last attribute in particular, Apperly and Butterfill say that this kind of belief representation is heavily dependent on executive cognitive resources.

Taking this into account, they specify limits on the nature of registration that differentiate it from belief. Specifically, they say that registration doesn't incorporate psychological role, abductive inference, normativity, or propositional content (p. 960 & 963). This helps them to specify signature limits on infant mindreading that can be

used as a basis for testing the theory. One such signature limit is an inability to understand how beliefs combine with each other and with other psychological states in guiding inferences and actions. Another signature limit is that, because infants are not representing propositional content, they should be unable to take mode of presentation into account, i.e. the way that an object appears to or is represented by the agent.

With respect to how the two systems are related, Apperly and Butterfill (2009) recognize that there is a spectrum of possibilities (p. 964). At one extreme, there may be no direct information flow between the two systems. The efficient system might influence attention and action, but would not directly provide input to explicit judgments about belief. At the other extreme, the flexible system might depend on the efficient system. But Apperly and Butterfill lean towards the former pole based on theoretical considerations and evidence that there is a marked dissociation between infant mindreading responses assayed using 'implicit' and 'explicit' measures.[7] One of the key findings from research employing the explicit version of the false belief task is that children who fail tend to say that Sally will look in the box that the doll is currently in (Wellman et al. 2001). Using a betting paradigm, Ruffman et al. (2001) found that children were confident in these responses. Apperly and Butterfill interpret this as indicating that the efficient mindreading system does not communicate directly with the flexible system (2009, p. 964).

Their theoretical reasoning is that cognitive efficiency depends on limitations on the complexity of input and processing. Accordingly, the efficient system could not provide belief information for the explicit system in most cases because the kinds of beliefs processed by the latter are generally too complex to be represented by the simple system (p. 964). Apperly (2011) gives a more specific characterization, suggesting that efficient mindreading processes will be fast and inflexible, will make low demands on general processing resources, will tend to resist strategic control, and may operate outside of awareness (p. 134).

[7] As detailed above, we think these labels are misleading.

5 A cooperative multi-system architecture for mindreading

In footnote 14 Butterfill and Apperly make some important and problematic concessions (2013, p. 620). The text preceding this footnote proposes that the mindreading principles they have presented can be extended to address a wider range of mindreading phenomena, such as desires. In the footnote they note that there is evidence that infants can do more than just track beliefs about the location of objects. In particular, they note evidence that infants can use communication to track belief (the Song et al. 2008 study described above), they can solve false belief tasks involving unexpected contents (He et al., 2011), and they can take into account inferences that an agent may make (Scott et al., 2010). In effect, Butterfill and Apperly acknowledge here that their account is unable to accommodate the extant evidence concerning infant mindreading. This raises the question of how their account should be construed, since it is, strictly, by their own admission falsified. The most natural interpretation is that they are sketching the initial foundations of an approach to theorizing the representation of beliefs and other states by infants and non-human animals – an approach whose central virtue is that it doesn't explicitly or implicitly presuppose a fully developed intentional folk psychology. In other words, they are not yet aiming for full empirical adequacy. While this is a reasonable theoretical approach, it is equally reasonable to press the account concerning these empirical limitations.

The crucial question is whether the account can be extended in a straightforward way to accommodate the unexplained evidence, or whether a comprehensive theory will require qualitatively different explanatory resources, or perhaps even a fundamentally different starting point. The most problematic aspect of Butterfill and Apperly's restricted empirical scope is that the overall pattern of evidence raises issues that their account does not appear well-equipped to deal with. Taken as a whole, the evidence reveals striking flexibility in infant belief representation abilities, whereas the mindreading principles described above are notably inflexible. It is not obvious how they can be straightforwardly extended to encompass the representation of beliefs that involved diverse properties, beliefs acquired through communication, and so on. We'll argue that explaining this flexibility requires an approach that is in key respects fundamentally different. A cooperative multi-system architecture is better able to explain infant belief representation than a parallel architecture, and causal

representation, schemas and models provide a more promising basis for flexible belief representation than does a rule-based approach of the kind described by Butterfill and Apperly.

5.1 The inflexibility of Butterfill and Apperly's representational scheme for efficient mindreading

In carefully seeking to avoid adult-like psychological concepts, Butterfill and Apperly have presented a very sparse toolkit of representational resources. The account has only minimal articulation of key features of intentional agency, incorporates no causal information, and has an inferential structure that consists of a small number of 'hard coded' or fixed rules. These features make it inherently poorly suited to explaining the flexibility of infant belief representation. It is certainly possible to elaborate the account to include other parameters of agency and a larger number of rules, but the account faces the same kind of problem as the behavior rules approach. That is, it requires a pre-specified rule for each significant variation, and the problem of combinatorial explosion ensures that, beyond a small number of variables, the number of rules required is implausible.

Some of the problems with the minimal articulation of the scheme can be appreciated by considering Butterfill and Apperly's account of the representation of goals, expressed in the presentation of the first principle. They appear to claim here that goals are inferred from patterns of movement, and are represented as the culmination of movement. But this is at odds with evidence that infants flexibly take into account environmental constraints in attributing goals, not just the movement itself (Gergely et al., 2002; Csibra et al., 2003). It also conflicts with evidence that infants can take into account the nature of the agent in attributing goals (Saxe et al., 2005), and can assign goals to agents, not just actions (Rochat et al., 2004). Indeed, Luo (2011) reported evidence that 10-month olds attribute the goal of grasping a *particular* object to an agent (and not some other object) if they attribute a preference for that object to the agent. Crucially, they expect the agent to have a preference for that object if she has previously grasped that particular object – but not if in grasping for the object she had a (true or false) belief that it was the only available object on the scene, in which case infants do not attribute a preference for the object. This study not only illustrates the flexibility with which infants identify and attribute goals but, more generally,

demonstrates that infants can flexibly integrate representations of agents' goals with representations of beliefs (that the object was or was not the only available object) observed behaviors (in this case a grasping action) and environmental opportunities (the presence of one or multiple graspable objects). Taken together, the evidence from these studies suggests that infants represent goals in a way that has more articulation than Butterfill and Apperly recognize, including flexible sensitivity to multiple aspects of the situation, action, and agent.

This shows that their theory is incomplete, but the lack of an articulated account of goal attribution causes deeper problems which undercut the theory's ability to explain the phenomena to which it is addressed, such as the ability of infants to predict Sally's actions in the false belief task. Specifically, the account doesn't provide a basis for assigning the goal of obtaining the doll to Sally's actions in the false belief condition. To appreciate this problem, first note that if goal attribution is based only on the immediate movement pattern then the infant should assign to Sally's actions the goal of approaching the box and engaging in movements involving the box. In order to assign to Sally's actions the goal of obtaining the doll the infant must take into account Sally's previous experience. More specifically, it is only by taking into account that Sally has registered the location of the doll as being in the box that the infant can assign to the action the goal of obtaining the doll. This sits awkwardly with the way that the principles are presented. The third principle states that an agent must correctly register an object in order to successfully perform an action that has the object as its goal. This seems to presuppose that the determination of the goal of the action is separate from and prior to the determination of the effect of registration on the action. In other words, registration determines whether the action will be successful but doesn't determine what the goal of the action is.

The account thus needs modification to explicitly recognize that registration can contribute directly to goal attribution. In itself this may seem a minor issue, but one of Butterfill and Apperly's core motivations is to avoid the inferential holism that they regard as a distinctive mark of mature mindreading performed by the flexible system. Recall that the first principle is supposed to allow the individual to infer goals without taking into account intentions, beliefs and other propositional attitudes. But if registration contributes to goal attribution then the individual is in effect taking beliefs

into account. Note further that one of the proposed signature limits of minimal mindreading is an inability to take into account interrelations amongst psychological states in guiding inferences and actions. But if registration contributes to goal attribution then the individual is, in effect, taking into account interrelations amongst psychological states. It thus begins to appear that infant belief representation also exhibits a degree of inferential holism.

The problem is worse than this, however, because taking into account immediate action together with registration also isn't sufficient for determining that the goal of Sally's action is obtaining the doll. When Sally was previously in the room she registered all of the objects that fell within her field, including the box that she placed the doll in. Accordingly, based only on Sally's immediate movements and her registrations, the infant has as much basis for attributing to Sally's action the goal of engaging with the box as obtaining the doll. In order to attribute the goal of obtaining the doll the infant needs some way of selecting amongst the various registrations Sally has made in the room. The most obvious mechanism for this is the attribution to Sally of an enduring motivational attitude towards the doll, such as an interest in or desire for it. Introducing motivation, however, brings in even greater inferential holism: there is mutual influence between attributions of motivational state, goals and registrations. Indeed, it looks like a full account of how the infant attributes the goal of obtaining the doll requires postulating mutual influence amongst representations of situation, motivation, actions, goals and registrations. This is a fairly rich inferential holism which violates the restrictions that Butterfill and Apperly have associated with minimal mindreading.

Thus, the immediate problem for Butterfill and Apperly's account is that it doesn't explain how the infant comes to attribute the goal of obtaining the doll in the false belief condition, and, as a result, the account doesn't succeed in explaining how the infant successfully predicts Sally's action. The larger problem is that it looks as though an elaborated explanation of goal attribution will fundamentally change the character of the account, introducing the kind of inferential holism that Butterfill and Apperly have been trying to avoid. And if goal attribution incorporates information about Sally's past experiences, motivation, and actions, then it looks like goal attribution forms an integral part of the tracking of Sally as an agent. They are *Sally's*

goals, not simply the endpoints of the actions that Sally engages in. Motivations, goals, beliefs and actions are all interrelated aspects of Sally as an integrated agent.

Another kind of problem for the account is that it will be difficult to accommodate these interdependencies with the kinds of representations that Butterfill and Apperly have proposed. The account has no mechanisms for generating new kinds of inferences 'on the fly', so every inference involved in belief representation must be 'hard coded' as a pre-specified rule. It thus confronts the same kind of combinatorial explosion problem that faces the behavior rules approach: as the number of variables increases linearly, the potential relations increase exponentially.

In section 3.1 we argued that generativity provides a solution to this type of problem, and we argued that causal representation provides an important form of generativity. In light of this it is significant that Butterfill and Apperly have deliberately evacuated all causal information from the minimal mindreading system. In the ordinary folk psychological understanding, perception has causal properties: it provides information for the agent which the agent can then use in various ways. The information-providing function of perception is linked to the functions of other states and activities of the agent. If the agent lacks information, she may engage in information-seeking activities such as directed attention or searching activities. This kind of structure is absent from Butterfill and Apperly's account. The second, third and fourth principles provide inferential rules that connect encountering and registration to action, but these rules are not represented as stemming from the causal properties of fields, encounterings and registrations. They are brute stipulations. The agent's field is simply an area extending from the agent, approximating the visual field but including no causal notion of vision or perception as information-producing. Encountering and registration approximate the information-providing function of perception, but they are defined in a way that includes no causal connection between perception and action. Encountering and registration are simply specified as necessary conditions for successful goal-directed action. This is very different and much weaker than representing perception as providing information *for* action. The latter is causal, the former is not.

Some clarification is needed here, given that Butterfill and Apperly use language which suggests that registration is causal. They frame their account in relation to a hypothetical individual called Lucky, and the principles of minimal mindreading are intended characterize a minimal set of representational resources that would allow Lucky to track beliefs in a restricted set of circumstances without requiring that she represent beliefs as such. In relation to the fourth principle, they say that:

> So far Lucky thinks of correct registration as a condition for the success of goal-directed action. This does not tell her anything about what happens if the condition is not met. In particular it tells her nothing about how an agent will act when she registers an object incorrectly. The fourth principle involves a switch from thinking of registration as a success condition *to thinking of it as a causal factor*. (2013, p. 619, emphasis added).

We have to treat this description as metaphorical, however, because minimal mindreading is supposed to be automatic and sub-personal. It is performed by a strongly encapsulated system that is largely independent of executive cognition and does not draw on general semantic memory. Lucky thus does not literally *think about* registration, or think of registration as playing a causal role in action. Nor could she, because registration, as it is defined, has no causal properties.

The fact that 'field', 'encountering' and 'registration' are not represented causally creates difficulties for any efforts to expand the account to represent a wider range of phenomena. For instance, if motivation and attention are introduced as variables, then rules will be needed that specify their interactions with all of the other variables, and these relations threaten to be exceedingly complex. For instance, Butterfill and Apperly's account allows an infant to predict that a goal-directed action aimed at an object will fail if the registration of the location of the object is incorrect. But it doesn't allow the infant to expect that the agent will search for the object when it is not re-encountered in the expected location. In the absence of the representation of causal interconnections between perception, motivation, goals and actions, the only way to predict searching would be by means of a 'hard coded' reasonless rule, and searching itself would be functionally opaque.

More generally, it is the reasonless nature of the inferences that forces the account to rely on a stock of predetermined rules, and it is the non-causal nature of the representations which ensures that the inferences are reasonless. Fundamentally, then, the non-causal nature of the representation makes it inherently inflexible, which in turn makes it poorly suited to account for the flexibility revealed in recent infant mindreading experiments. For example, in the false belief condition of the Southgate et al. (2010) experiment described in section two, Sally points to one of the boxes, names the toy in the box 'sefo', and asks the infant to retrieve the sefo. Despite Sally's pointing, the infant goes to the other box and retrieves for Sally the toy inside. This requires the infant to connect Sally's false belief with a referential intention in a highly idiosyncratic scenario. A large stock of rules can give the appearance of flexibility, but evidence for flexibility in idiosyncratic situations makes this kind of explanation unlikely because of the combinatorial explosion problem. Evidence for a capacity to flexibly interpret and respond to complex idiosyncratic situations points to generative capacities for representation and action.

As has been noted, Butterfill and Apperly have tried to avoid causal representations of psychological states and inferential holism out of concern that this would involve attributing to infants a 'full-blown' adult-like folk psychology. We agree with them that attributing an adult-like folk psychology to infants is undesirable, but nevertheless believe that their motivation is misguided. This is because there are theoretical ways to capture this kind of flexible, inferential holism that don't require positing a full, adult-like folk psychology.

5.2 The advantages of a cooperative multi-system architecture

Before elaborating on the representational basis for holistic inferential interpretation of agency, we need to return to the issue of the overall cognitive architecture of mindreading. As described in section 4, Apperly and Butterfill (2009) use an analysis of the tradeoff between efficiency and flexibility to argue that the early-developing efficient mindreading system should take as input only limited forms of information, perform simple processing, and operate largely in parallel with the flexible mindreading system. We'll now challenge this line of reasoning.

As a first step it will help to consider the architecture of mindreading in relation to a larger picture of cognitive architecture. One of the most fundamental and sustained patterns of evidence to emerge from neuroscience research is that there are many distinct neural regions that perform specialized information processing functions (Gazzaniga et al., 2002, ch. 1). But in addition to local specialization there is also evidence that the central nervous system possesses an integrative hierarchical macro-architecture (Fuster, 2004; 2008). With respect to cortical organization, the primary sensory and motor areas form the base of the hierarchy, while downstream areas successively integrate information more widely and perform more complex cognitive functions. The rostral prefrontal cortex and the polymodal associative sensory cortices form the highest level of the hierarchy.

This progressively integrative hierarchy differs significantly from the view of cognitive architecture implied by Apperly and Butterfill's analysis of the efficiency-flexibility tradeoff. Specifically, the executive isn't the only system that integrates information widely: intermediate systems exhibit various kinds of specialization, but they also integrate relatively wide-ranging information, perform information processing functions that are complex and flexible, and engage in rich interactions with other systems. We can illustrate this with two examples: the orbitofrontal cortex (OFC) and the mid-level visual system.

The OFC plays a specialized role in determining the incentive value of action outcomes (Schoenbaum et al., 2003). It receives input from many sensory areas, has connections with motor systems, and is reciprocally connected with the basolateral amygdala and the parahippocampal region. In a review, Schoenbaum et al. (2003) present evidence that the OFC integrates sensory information with abstract relational processing to determine the value of outcomes, and interacts flexibly with other systems depending on the task requirements. Interactions with the basolateral amygdala help to determine the incentive values of outcomes: the amygdala and the OFC both process value information, but the amygdala processes relatively simple information concerning the value of the current outcome, while the OFC processes more complex relational information used to guide the actions, including sensory cues associated with outcomes. The OFC interacts with the parahippocampal region in tasks that require rapid, flexible abstraction, as is the case in a non-match to sample

task where the correct choice from a pair of items is the one that doesn't match the sample item.

The OFC thus shows that a specialized system can exhibit complex, flexible information processing that draws on wide-ranging sources of information. The example also reveals a way that rich interactivity can enhance efficiency. A strongly encapsulated system must perform all of the information processing required for producing solutions to the problems it solves internally, whereas a richly interacting system can rely on other systems for critical information. The construction of flexible assemblages for task performance allows for complex information processing tailored to the shifting demands of different tasks.

The 'mid-level visual system' illustrates the fact that specialised systems can generate rich information that is consciously accessible. The information represented by this system is primarily spatiotemporal, and the system supports spatiotemporal inferences concerning objects, such as that an object continues to exist when it disappears behind a barrier, and, if it has an appropriate trajectory, that it will reappear at a particular time and location (Kahneman et al., 1992; Carey, 2009). Under certain circumstances the mid-level visual system is relatively insensitive to non-spatiotemporal information, and will interpret a stimulus sequence as depicting a single continuous object, even though this involves a transformation from one type of object to another, such as a rabbit transforming into a bird (Carey, 2009, pp. 72-3). Nevertheless, the mid-level visual system plays a powerful and relatively generalized role in object representation. It supports fairly rich inferences about objects, and is capable of learning complex relations, such as the fact that unsupported objects tend to fall (Baillargeon, 1998).

Two further examples will help to show the benefits of cooperative relations between conscious cognitive control and lower order systems. Research with the patient HM played a foundational role in the development of the view that there are multiple memory systems, including a conscious declarative system and a non-conscious procedural system involved in motor skill learning. Milner (1962) trained HM to perform a mirror drawing task, and despite the fact that HM had no conscious memories of prior learning his ability improved with increased practice. This has been

taken to show that motor skills do not depend on conscious control, but as Stanley and Krakauer (2013) point out, HM required explicit verbal instructions each time he performed the task. This highlights one of the roles that conscious cognition plays in normal skilled action: it controls the circumstances in which the skill is exercised and how it is exercised. In other words, while it is true that non-conscious systems make a substantial contribution to motor skill, it is not the case that these skills are entirely 'procedural'. Higher order conscious knowledge plays a crucial role in the exercise of motor skills.[8]

A final example reinforces the previous point and helps to illuminate the integrative role that higher cognition can play in cooperative relations with lower order systems. Language comprehension is a multi-level process that includes automatic components, higher level interpretive processes, and higher level influences on lower order processing. The interference found in the classic Stroop task (Posner and Snyder, 1975) shows that when attention is directed to a word, reading the word is an automatic, obligatory process, occurring even when it is irrelevant to the task requirements (naming the color of the text). But reading sentences and passages is a voluntary process, depending on sustained conscious attention. Comprehension involves the construction of a situation model (Zwaan and Radvansky, 1998), which provides an integrated interpretation. Lashley (1951) illustrated the importance of this integration with examples that reveal the role of high level interpretation for resolving lower level ambiguities such as occur with homophones and homographs. For instance, when the following sentence is spoken, the homophone ˈrītiNG must be resolved from the context: "Rapid righting with his uninjured hand saved from loss the contents of the capsized canoe" (p. 120). Moreover, the correct interpretation only becomes apparent when the whole sentence is understood, because the local context ("with his uninjured hand") is misleading. This shows that in the *sound → word → meaning* processing stream, high level processing of meaning influences *sound → word* processing.

In sum, these examples indicate that Apperly and Butterfill's analysis of the efficiency-flexibility trade-off is mistaken; the demands of efficiency don't require

[8] Christensen et al. (submitted) argue that conscious cognitive control plays an even stronger role in skilled action than suggested by Stanley and Krakauer, but we will not pursue this issue here.

strong encapsulation and simple information processing. Specialised systems can integrate information widely and engage in rich interactions with other specialised systems and with executive cognition. Indeed, this is an efficient form of organization for performing complex cognitive tasks because the processing burden is distributed and varied information processing resources can be accessed as required by shifting task demands.

5.4 The integrative flexibility of infant mindreading suggests a cooperative multi-system architecture

The considerations raised in the previous section give background plausibility to the view that mindreading involves a cooperative multi-system architecture. But we can find more direct evidence for this within the mindreading literature.

One of the most basic forms of evidence is inherent in the nature of the standard false belief task: generating the right prediction about Sally's action in the false belief condition requires integrating diverse forms of situational information over an extended period of time, which means that the information processing is dependent on working memory and the construction of an integrated situation model. A strongly encapsulated non-conscious system is not capable of this kind of information processing. One possibility that must be considered is that there is a non-conscious mindreading system whose operation is parasitic on executive cognition. That is, executive cognition is responsible for the construction of the integrated situation model that connects Sally's initial placing of the doll in the box, Anne's moving the doll in Sally's absence, and Sally's subsequent approach to the original box. This information serves as input for the non-conscious belief representation system, but this system does not in turn provide output that feeds back into the integrated situation model. We can rule this possibility out, however. The fact that infants attribute to Sally the goal of obtaining the doll shows that their representation of Sally's belief concerning the doll's location does become incorporated into the integrated situation model.

The findings from the verbal communication, active helping and naming paradigms provide especially compelling evidence that belief information is incorporated into the infant's integrated situation model. When the confederate tells Sally the new

location of the doll (Song et al. 2008), the infant must be able to directly relate this declarative information to her representation of Sally's belief about the doll's location if she is to revise her expectations concerning what Sally will do. If an encapsulated non-conscious belief system were able to track beliefs independently of executive cognition (contrary to the argument of the preceding paragraph) it should be insensitive to changes in declarative knowledge. In the case of the active helping experiment (Buttelmann et al. 2009), the infant must use her representation of Sally's belief in order to interpret Sally's actions (attempting to open the box), and then use this information to inform the infant's own action, which is to retrieve for Sally the doll from the other box. In the case where Sally points to one of the boxes and labels the toy inside a sefo (Southgate et al. 2010), the infant must again use her representation of Sally's belief in order to interpret Sally's referential intention. The infant must then use this information to inform her own action, which is to retrieve for Sally the toy in the other box.

Taken as a whole, these experiments provide compelling evidence that infants are flexibly integrating belief information with other information in an integrated situation model. This implies that belief information is being processed by the executive system, which incorporates the flexible cortical-hippocampal memory system (Eichenbaum 2000). On the other hand, the evidence for sensitivity to various features of agency and belief at 3-7 months (Woodward, 1998; Sommerville et al., 2005; Kovacs et al. 2010; Southgate et al., 2014) suggests that the representation of agency involves adaptively specialized systems. These two lines of evidence can be reconciled by the view that mindreading involves a cooperative multi-system architecture.

5.5 A framework for flexible belief representation

This returns us to the problem of explaining how infants can be capable of holistically inferential agency representation without attributing to them a 'full blown' adult-like folk psychology. Some potential ingredients for an account were introduced in our discussion of the behavior rules approach: statistical representation, causal representation, schemas, and models. We'll now sketch a way of combining these ingredients.

We won't speculate about the specific processing characteristics of the early-developing adaptively specialized agency systems, but at minimum they are likely to bias attention to key features of agency. Statistical and causal learning will generate increasingly differentiated representations of these features and relations amongst them. Statistical learning is known to play a powerful role in the segmentation of speech sounds into words (Saffran et al., 2001), and is likely to play a similar role in segmenting core agentic structure. Causal learning will similarly help to differentiate key features and structures, but as well as pattern it identifies causal properties. The diagnosticity of causal learning will help to focus attention and build patterns of salience. And developing causal knowledge will build flexible inferential connections between the various aspects of agency, such as motivational states, perceptual awareness, goals, and actions. In combination, statistical and causal learning will build an increasingly rich and articulated stock of schemas for various aspects of agency, and learning will occur at multiple levels, including low level structures for basic agentic parsing and higher level schemas and causal representations for situations and integrated action structures. The stock of schemas and causal knowledge provides the basis for the construction of models for particular situations, and model construction becomes faster, more flexible and more detailed as this representational base becomes richer. Situation models are integrated: they are constructed progressively, and the elements are mutually constraining.

This representational system is generative, and gains increasing power and flexibility as it gains greater articulation. It supports inferential holism with a suite of interconnected mechanisms, none of which presuppose the very rich, highly articulated psychological conceptions involved in adult mindreading. In this respect it's important to note that an infant can have very simple interconnected representations of causal features of agency without having abstract conceptualizations of agency. It's also important to note that this representational system supports holistic inference without requiring that holistic inference be based on effortful abductive inference (the cognitive mechanism that Apperly and Butterfill associate with the flexibility of adult mindreading). Model construction is a holistic inferential process, but when it is scaffolded by a rich stock of schemas and causal knowledge it can occur rapidly and with little effort. Text comprehension is a useful reference point here: it can be very cognitively demanding when the material is

complex, difficult, and perhaps unusual (e.g., Joyce's Ulysses), but can also be relatively effortless when the text is simple and the subject matter familiar (e.g., a story in the local newspaper reporting that the beaches in the area have again received a poor rating in pollution tests).

Suggestive evidence in support of this account comes from findings that causal learning plays a role in infants' representations of agency. Thus, Sommerville et al. (2005) gave 3-month-old infants experience picking up objects with 'sticky mittens' that had Velcro fabric on the palms. They found that this experience allowed the infants to interpret the reaching actions of an agent with a similar mitten as goal-directed. Sommerville and Woodward (2005) found that 12-month-old infants understood the causal structure of an action in which a toy was moved by pulling a cloth the toy was sitting on. These results show that causal learning contributes to the representation of agency from a very young age, and in light of the considerations we've discussed, this may be an important clue to the basis of the substantial inferential flexibility shown by older infants in representations of agency.

Our account also provides a reinterpretation of evidence that Apperly and Butterfill view as supporting the claim that efficient mindreading occurs automatically and independently of executive control. Samson et al. (2010) employed a task in which participants viewed an image of a room in which an avatar is standing in the middle and facing either to the left or to the right. On each trial, anywhere from zero to three red discs are displayed on the walls of the room, and due to the varying positions of the discs the avatar can see either none, some, or all of them. In the condition where participants simply had to report how many discs they themselves see (irrespective of the perspective of the avatar), it was found that performance was impaired when the participant could see a different number of discs to the avatar. Samson et al. interpreted this as showing that the participants were automatically calculating the number of discs the avatar could see, even though this was irrelevant to the task.

While we agree that the results show that others' perspectives are in some circumstances calculated spontaneously and efficiently, it is important to be cautious about interpreting this as evidence of a dedicated mindreading system. It is also possible, for example, that the effect is at least partially driven by domain-general

spatial-cueing, with the avatars serving as cues to trigger attention either to the left or to the right. In fact, this is the interpretation offered by Santiesteban et al. (2014), who replicated Samson et al.'s effect using arrows instead of avatars. What our multi-systems account suggests, however, is that the disagreement between Santiesteban et al.'s *domain-general* interpretation and Samson et al.'s *domain-specific* interpretation may reflect a false dichotomy: the paradigm likely involves both domain-specific mechanisms for identifying agentic features (a human-like body, the gaze direction of the avatar), and a domain-general attentional mechanism (spatial cueing) that is engaged by the identification of agentic features. Indeed, the view that this paradigm engages a medley of domain-specific and domain-general systems is supported by earlier results showing that gaze cueing is not only similar to spatial cueing using arrows, but also, and in important ways, different from it: gaze cueing, unlike spatial cueing with arrows, is automatic in the sense that faces (but not arrows) trigger spatial cueing even if the gaze direction of the face has very low cue validity, and participants are informed of this (Driver et al.1999); and participants tend to evaluate faces in quasi-moral terms (i.e. as trustworthy or untrustworthy) depending on the cue validity of their gaze direction (Bayliss and Tipper 2006). In order to develop and test our multi-systems interpretation of these findings further, it would be important to investigate similarities and differences between avatars and other types of cue (such as arrows) in this paradigm. For example, unlike Santiesteban et al.'s account, our account generates the prediction that the validity of the avatar's gaze direction may have little effect upon participants' performance. On the other hand, unlike Samson et al.'s (and Butterfill and Apperly's) automatic mindreading account, our account predicts that various other systems can cooperate with the systems engaged in this pararidgm. Thus, we predict that performance on the Samson paradigm could be modulated by manipulating participants' beliefs about the avatar, for example about whether the avatar is sighted or blind, whether a pair of goggles that s/he is wearing is transparent or opaque, etc. Intriguingly, this latter prediction is motivated by the findings of Teufel et al. (2009), who reported that participants' processing of gaze direction was facilitated when a subject believed that a person wearing goggles was able to see through them (as opposed to the goggles being opaque).

More generally speaking, the experiment has similarities to the classic Stroop task, and the underlying mechanisms may also have structural similarities. As we discussed

above, when attention is directed to a word, the word is read automatically. This doesn't show that word reading is performed by an encapsulated non-conscious system that operates independently of executive control, however. Word reading feeds into higher order voluntary reading processes, and is itself under executive control through the control of attention and eye gaze. On our account, high level mindreading is supported by low level agency parsing mechanisms that function analogously to the mechanisms underlying word reading. When an individual is viewing or otherwise becoming acquainted with a situation, these mechanisms will automatically detect and process low level agentic structure in the situation. But these mechanisms do not operate in parallel with executive cognition; they feed into it and are in turn modulated by higher order control.

But perhaps the most pressing problem for our account is the core puzzle introduced by the evidence for infant belief representation: explaining why infants succeed on the 'implicit' versions of the false belief task but fail on 'explicit' versions. While we will not attempt to provide a comprehensive answer to this here, we do want to highlight the resources that our account offers for addressing the problem. A key initial point to note is that learning organizes knowledge for the particular kinds of activities and problems the individual engages in. Schemas and other forms of knowledge organization help to make relevant knowledge differentially available in familiar situations. The converse of this is that the individual may possess the underlying knowledge required to solve a particular problem, but not be able to deploy this knowledge because she lacks the relational understanding required to apply the knowledge in the situation. Maier's (1931) pendulum experiments illustrate this kind of case. The participants undoubtedly possessed the concept 'pendulum', but could not relate the concept to the situation.

In the case of mindreading, the early organization of the representation of agency in infants will be primarily in support of their own situational interpretations and activities. Being able to deploy knowledge in response to questions is a very different kind of cognitive problem than using knowledge in the course of immersed situational interpretation and action. In order to retrieve information from long term memory it is necessary to construct retrieval cues that access the information, and infants may lack the knowledge organization required to formulate the appropriate retrieval cues in

response to questions. In other words, they lack the appropriate cognitive 'hooks' to link questions to the relevant knowledge.

But the way we have just described this makes it seem as if the problem is only one of accessing existing knowledge. In fact, we believe that the infant's representations of agency will undergo a qualitative shift as they become increasingly articulated. This brings us to a final point of difference with Apperly and Butterfill's account. They identify two kinds of belief representation: the non-conceptual form of representation performed by their minimal mindreading system, and the adult-like conceptual representation of beliefs 'as such'. In contrast, while we acknowledge that belief reasoning is likely to be supported by various distinct forms of representation (e.g. statistical representations, causal representations, schemas, models and linguistically encoded representations), we deny that any neat division into two representational systems is motivated by existing data or by theoretical considerations, and we also maintain that cooperation of distinct representational systems, not segregation, is the more likely default.

Having said this, it is important also to acknowledge that the development of increasingly explicit conceptualizations of beliefs and other aspects of agency will mark a qualitative shift in belief representation. Amongst other benefits, explicit conceptualizations provide the cognitive 'hooks' that allow knowledge to be accessed and processed in complex reasoning and in response to questions. At the same time, however, the development of explicit conceptualizations of belief and other features of agency is only one part of a larger suite of representational changes that support skilled social interaction. One of the core ideas informing our account is that the representation of belief is part of a larger suite of abilities involved in the representation of situations. Increasingly sophisticated understanding of various kinds of social situations makes a large contribution to the ability to represent beliefs (Michael et al., 2013). In this respect we are echoing Apperly (2011), who likewise highlights the role of social knowledge and situation models in mindreading (especially pp. 128-132). Here our main differences are in emphasis. Apperly recognizes that social knowledge can facilitate belief representation, but his prevailing focus is on the fact that social knowledge can reduce the need for accurate belief representation. We don't disagree that this is one of the major cognitive effects of rich

social knowledge, but we feel that Apperly doesn't fully capture the enabling role of social knowledge in belief representation. Highly articulated social knowledge provides a high degree of holistic inferential flexibility and allows subtle and precise belief representation.

6 Conclusions

It is worth distinguishing amongst the various claims we have made because it is possible to accept some without accepting the whole package. First of all, we have argued that the behavior reading approach does not do justice do the many findings that have been reported from various paradigms. Most importantly in this context, we have emphasized the significance of generativity and diagnosticity for understanding complex abilities that exhibit considerable flexibility. More specifically, we believe that it is likely that causal representation, schemas and models all contribute to belief reasoning, and that some of the key benefits they provide are generativity and diagnosticity. But perhaps the strongest consideration against the behavior reading approach is the substantial body of evidence that young infants are sensitive to many features of intentional agency and their interrelations (in addition to the literature cited above, for a review see Woodward et al. 2009).

With respect to nativism, we've suggested that the balance of evidence favors the view that the development of mindreading involves qualitative changes in representational ability rather than just changes in access to executive cognition and improved ability to apply a belief concept. But it would be valuable to differentiate more clearly patterns of development in mindreading, and agency representation more generally, that could be expected on the basis of a nativist account and patterns distinctively predicted by qualitative change theories. In this respect, on the basis of our account it can be expected that learning exhibits a constructivist pattern in which the development of schemas and causal representation improves the ability to construct integrated situation models, which in turn improves the articulation and richness of causal and schematic representations. As the ability to differentiate simple relations in intentional agency develops, this opens up the ability to differentiate more complex relations, which will result in new conceptualizations at higher levels.

With respect to Apperly and Butterfill's two systems theory, the strongest consideration against an encapsulated low level mindreading system lies in the nature of the false belief task: it requires the structured integration of diverse forms of information over extended periods of time. This is a signature of tasks that depend on executive cognition. Whether or not our specific claims concerning the architecture and representation of beliefs and agency are viewed as convincing, we think there are strong grounds for examining the possibility that mindreading involves a cooperative multi-system architecture of some kind.

References

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

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

Baillargeon, R. (1998). Infants' understanding of the physical world. In M. Sabourin, F. Craik, & M. Robert (Eds.), *Advances in psychological science, Vol. 2: Biological and cognitive aspects* (pp. 503–529). Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. doi:10.1016/j.tics.2009.12.006

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

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bayliss, A. P., & Tipper, S. P. (2006). Predictive Gaze Cues and Personality Judgments Should Eye Trust You? *Psychological Science*, *17*(6), 514–520. doi:10.1111/j.1467-9280.2006.01737.x

Birch, S. A. J., & Bloom, P. (2007). The Curse of Knowledge in Reasoning About False Beliefs. *Psychological Science*, *18*(5), 382–386. doi:10.1111/j.1467-9280.2007.01909.x

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. doi:10.1016/j.cognition.2009.05.006

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

Call, J., & Tomasello, M. (1999). A Nonverbal False Belief Task: The Performance of Children and Great Apes. *Child Development*, *70*(2), 381–395. doi:10.1111/1467-8624.00028

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192. doi:10.1016/j.tics.2008.02.010

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. doi:10.1111/1467-8624.00333

Carruthers, P. (2013). Mindreading in Infancy. *Mind & Language*, *28*(2), 141–172. doi:10.1111/mila.12014

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395. doi:10.1016/0885-2014(94)90012-4

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133. doi:10.1016/S0364-0213(02)00112-X

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze Perception Triggers Reflexive Visuospatial Orienting. *Visual Cognition*, *6*(5), 509–540. doi:10.1080/135062899394920

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). New York, NY, US: Oxford University Press.

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, *1*(1), 41–50. doi:10.1038/35036213

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, *8*(4), 143–145. doi:10.1016/j.tics.2004.02.004

Fuster, J. M. (2008). *The prefrontal cortex* (4th ed.). London: Elsevier.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2002). *Cognitive Neuroscience: The Biology of the Mind*. Norton.

Gergely, G., Bekkering, H., & Király, I. (2002). Developmental psychology: Rational imitation in preverbal infants. *Nature*, *415*(6873), 755–755. doi:10.1038/415755a

Gureckis, T. M., Goldstone, R. L., & Hogan, P. C. (2010). Schema. In *The Cambridge Encyclopedia of the Language Sciences*. Cambridge, England: Cambridge University Press. Retrieved from http://cognitrn.psych.indiana.edu/rgoldsto/pdfs/schemaforlanguage.pdf

Harris, P. L. (2005). Conversation, Pretense, and Theory of Mind. In J. W. Astington & J. A. Baird (Eds.), *Why language matters for theory of mind* (pp. 70–83). New York, NY, US: Oxford University Press.

He, Z., Bolz, M., & Baillargeon, R. (2011). False-belief understanding in 2.5-year-olds: evidence from violation-of-expectation change-of-location and unexpected-contents tasks. *Developmental Science*, *14*(2), 292–305. doi:10.1111/j.1467-7687.2010.00980.x

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219. doi:10.1016/0010-0285(92)90007-O

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. doi:10.1037/0033-295X.85.5.363

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science*, *330*(6012), 1830–1834. doi:10.1126/science.1190792

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–131). New York: Wiley.

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, *9*(10), 459–462. doi:10.1016/j.tics.2005.08.002

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind." *Trends in Cognitive Sciences*, *8*(12), 528–533. doi:10.1016/j.tics.2004.10.001

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition*, *121*(3), 289–298. doi:10.1016/j.cognition.2011.07.011

Maier, N. R. F. (1931). Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, *12*(2), 181–194. doi:10.1037/h0071361

Michael, J., Christensen, W., & Overgaard, S. (2013). Mindreading as social expertise. *Synthese*, 1–24. doi:10.1007/s11229-013-0295-z

Milner, B. (1962). Les troubles de la memoire accompagnant des lesions hippocampiques bilaterales. *Physiologie de L'hippocampe*, 257–72.

Morgan, C. L. (1903). *An introduction to comparative psychology* (2nd ed.). London: W. Scott.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*, *308*(5719), 255–258. doi:10.1126/science.1107621

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Perner, J., & Ruffman, T. (2005). Infants' Insight into the Mind: How Deep? *Science*, *308*(5719), 214–216. doi:10.1126/science.1111656

Piaget, J. (1952). *The origins of intelligence in children*. (M. Cook, Trans.). New York, NY, US: W W Norton & Co.

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–82). Hillsdale, NJ: Erlbaum.

Rochat, P., Striano, T., & Morgan, R. (2004). Who is doing what to whom? Young infants' developing sense of social causality in animated displays. *Perception*, *33*, 355–369.

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review*, *34*(3), 265–293. doi:10.1016/j.dr.2014.04.001

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does Eye Gaze Indicate Implicit Knowledge of False Belief? Charting Transitions in Knowledge. *Journal of Experimental Child Psychology*, *80*(3), 201–224. doi:10.1006/jecp.2001.2633

Sabbagh, M. A., Moses, L. J., & Shiverick, S. (2006). Executive Functioning and Preschoolers' Understanding of False Beliefs, False Photographs, and False Signs. *Child Development*, *77*(4), 1034–1049. doi:10.1111/j.1467-8624.2006.00917.x

Saffran, J. R., Senghas, A., & Trueswell, J. C. (2001). The acquisition of language by children. *Proceedings of the National Academy of Sciences*, *98*(23), 12874–12875. doi:10.1073/pnas.231498898

Samson, D., Apperly, I. A., 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. doi:10.1037/a0018729

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), 929–937. doi:10.1037/a0035175

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret Agents Inferences About Hidden Causes by 10- and 12-Month-Old Infants. *Psychological Science*, *16*(12), 995–1001. doi:10.1111/j.1467-9280.2005.01649.x

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures* (1 edition.). Hillsdale, N.J.; New York: Psychology Press.

Schoenbaum, G., Setlow, B., & Ramus, S. J. (2003). A systems approach to orbitofrontal cortex function: recordings in rat orbitofrontal cortex reveal interactions with different learning systems. *Behavioural Brain Research*, *146*(1–2), 19–29. doi:10.1016/j.bbr.2003.09.013

Scott, R. M., & Baillargeon, R. (2009). Which Penguin Is This? Attributing False Beliefs About Object Identity at 18 Months. *Child Development*, *80*(4), 1172–1196. doi:10.1111/j.1467-8624.2009.01324.x

Scott, R. M., Baillargeon, R., Song, H., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, *61*(4), 366–395. doi:10.1016/j.cogpsych.2010.09.001

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind Eyes: An Absence of Spontaneous Theory of Mind in Asperger Syndrome. *Science*, *325*(5942), 883–885. doi:10.1126/science.1176170

Sommerville, J. A., & Woodward, A. L. (2005). Infants' Sensitivity to the Causal Features of Means–End Support Sequences in Action and Perception. *Infancy*, *8*(2), 119–145. doi:10.1207/s15327078in0802_2

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*(1), B1–B11. doi:10.1016/j.cognition.2004.07.004

Song, H., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition*, *109*(3), 295–315. doi:10.1016/j.cognition.2008.08.008

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*(6), 907–912. doi:10.1111/j.1467-7687.2009.00946.x

Southgate, V., Senju, A., & Csibra, G. (2007). Action Anticipation Through Attribution of False Belief by 2-Year-Olds. *Psychological Science*, *18*(7), 587–592. doi:10.1111/j.1467-9280.2007.01944.x

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, *130*(1), 1–10. doi:10.1016/j.cognition.2013.08.008

Stanley, J., & Krakauer, J. W. (2013). Motor Skill Depends on Knowledge of Facts. *Frontiers in Human Neuroscience*, *7*. Retrieved from http://www.frontiersin.org/Human_Neuroscience/abstract/55476

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of Beliefs by 13-Month-Old Infants. *Psychological Science*, *18*(7), 580–586. doi:10.1111/j.1467-9280.2007.01943.x

Teufel, C., Alexis, D. M., Todd, H., Lawrance-Owen, A. J., Clayton, N. S., & Davis, G. (2009). Social Cognition Modulates the Sensory Coding of Observed Gaze Direction. *Current Biology*, *19*(15), 1274–1277. doi:10.1016/j.cub.2009.05.069

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology*, *30*(1), 172–187. doi:10.1111/j.2044-835X.2011.02067.x

Träuble, B., Marinović, V., & Pauen, S. (2010). Early Theory of Mind Competencies: Do Infants Understand Others' Beliefs? *Infancy*, *15*(4), 434–444. doi:10.1111/j.1532-7078.2009.00025.x

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. doi:10.1111/1467-8624.00304

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. doi:10.1016/0010-0277(83)90004-5

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. doi:10.1016/S0010-0277(98)00058-4

Woodward, A. L. (2005). The infant origins of intentional understanding. In Robert V. Kail (Ed.), *Advances in Child Development and Behavior* (Vol. Volume 33, pp. 229–262). JAI. Retrieved from http://www.sciencedirect.com/science/article/pii/S0065240705800096

Woodward, A. L., Sommerville, J. A., Gerson, S., Henderson, A. M. E., & Buresh, J. (2009). The Emergence of Intention Attribution in Infancy. In Brian H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 51, pp. 187–222). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742109510067

Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology*, *30*(1), 156–171. doi:10.1111/j.2044-835X.2011.02060.x

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. doi:10.1037/0033-2909.123.2.162