FLEXIBLE GOAL ATTRIBUTION IN EARLY MINDREADING

Forthcoming in *Psychological Review*

John Michael
Department of Cognitive Science, Central European University, Budapest
michaelj@ceu.edu

&

Wayne Christensen
Department of Cognitive Science, Macquarie University, Sydney
wayne.christensen@gmail.com

ABSTRACT

The two systems theory developed by Apperly and Butterfill (2009; Butterfill & Apperly, 2013) is an influential approach to explaining the success of infants and young children on implicit false belief tasks. There is extensive empirical and theoretical work examining many aspects of this theory, but little attention has been paid to the way in which it characterizes goal attribution. We argue here that this aspect of the theory is inadequate. Butterfill and Apperly's characterization of goal attribution is designed to show how goals could be ascribed by infants without representing them as related to other psychological states, and the minimal mindreading system is supposed to operate without employing flexible semantic-executive cognitive processes. But research on infant goal attribution reveals that infants exhibit a high degree of situational awareness that is strongly suggestive of flexible semantic-executive cognitive processing, and infants appear moreover to be sensitive to interrelations between goals, preferences and beliefs. Further, close attention to the structure of implicit mindreading tasks – for which the theory was specifically designed – indicates that flexible goal attribution is required to succeed. We conclude by suggesting two approaches to resolving these problems.

Keywords: theory of mind, goal attribution, two systems theory, infancy, cognitive development, minimalism

One of the central challenges for contemporary developmental psychology is to provide a satisfactory explanation of the following, seemingly paradoxical, pattern of findings. On the one hand, children do not tend to succeed at explicit verbal false belief tasks until about four-and-a-half years of age (Wimmer & Perner, 1983; Wellman et al., 2001; but see Rubio-Fernández & Geurts, 2012). On the other hand, however, studies using implicit measures have now produced extensive evidence that infants are sensitive to others' false beliefs by the second year of life (Onishi & Baillargeon, 2005; Surian et al., 2007; for reviews, see Apperly, 2011, chap. 3; Baillargeon et al., 2010), and perhaps as early as the middle of their first year (Kovacs et al., 2010; Southgate et al., 2014).

The debate about how to account for this conflict has been structured by a contrast between rich and lean accounts. Rich accounts (e.g. Baillargeon et al., 2010; Kovacs et al., 2010; Southgate et al., 2007; Carruthers, 2013) maintain that infants represent others' beliefs by around one year or earlier, and then offer various explanations to account for the lag in performance on explicit verbal false belief tasks. In contrast, lean accounts (Perner & Ruffman, 2005; Heyes, 2014; Ruffman, 2014) deny that children represent beliefs before about four, and interpret infants' performance on implicit false belief tasks as resulting from behavior reading based on statistical learning. In between these two extremes, Apperly and Butterfill (2009; Apperly, 2011; Butterfill & Apperly, 2013) have proposed a *two-systems* account of mindreading. They explain the infant data by positing an early-emerging, simple, modular representational system that enables infants and young children to track beliefs in restricted circumstances, but does not afford the representation of beliefs 'as such'. The latter ability emerges when a second system for belief representation employing working memory and generalized, flexible long-term memory is in place. For convenience we'll refer to these two systems as the 'minimal' and 'flexible' mindreading systems, following Apperly and Butterfill's terminology.

Although Apperly and Butterfill's theory has received a great deal of attention, one crucial element has not yet been examined closely – the way in which the theory accounts for goal attribution in minimal mindreading. In what follows we show that there are problems with this aspect of the theory. We begin by briefly recapitulating Apperly and Butterfill's theory ('Minimal mindreading'). We then argue that the account of goal attribution is underspecified, and that it will be difficult to remedy this shortcoming, given that the overall theory by design eschews the cognitive flexibility that appears to be required to explain context-sensitive goal attribution ('Interpreting principle 1'). Next we summarize a selection of key empirical findings which indicate that infant goal attribution indeed shows the kind of context-sensitivity that is difficult for Apperly and Butterfill's theory to explain (Evidence against principle 1'). We go on to argue that this kind of flexible, context-sensitive goal attribution also plays a role in infant performance on false belief tasks ('Problems in applying principles 2-4'). Finally, we canvass options for revising the theory and suggest that the most plausible approach involves fundamental changes ('Two kinds of response to these problems').

Minimal Mindreading

Butterfill and Apperly (2013) (henceforth B&A) provide a functional specification of the early-emerging mindreading system, which was described only briefly in Apperly & Butterfill (2009). Their aim is to specify a set of principles and representations that could enable a system to track beliefs in simple situations without representing beliefs 'as such'. Because of the simplicity of the principles and representations employed by such a system, it would lack the flexibility that B&A associate with adult mindreading. One of the attractions of this approach is that it offers the potential to explain infants' success on implicit false belief tasks while simultaneously providing a basis for explaining why success on explicit versions of the

false belief task is delayed. The core of B&A's minimal mindreading system, then, is a set of four principles for reasoning about belief-like states of other agents on the basis of their behavior, which are as follows.

Principle 1: Bodily movements form units that are directed towards goals. As B&A put it:

> 'We stipulate that for an outcome, *g*, to be the goal of some bodily movements is for these bodily movements to occur in order to bring about *g*; that is, *g* is the function of this collection. Here 'function' should be understood teleologically. On the simplest teleological construal of function, for an action to have the function of bringing about *g* would be for actions of this type to have brought about *g* in the past and for this action to occur in part because of this fact' (2013, p. 613).

The virtue of this way of representing goals is that it allows them to be inferred from actions without appealing to intentions, beliefs, preferences or other psychological states (p. 613).

Principle 2: B&A introduce two kinds of representation – *fields* and *encounterings* – that together serve as a simplified surrogate for visual perception. B&A stipulate that a field is a 'set of objects' (p. 614) in an area specified in relation to the agent. The agent's field is determined by factors such as proximity, lighting, eye direction, opaque 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, then, is that the agent must encounter an object before she can engage in goal-directed actions aimed at the object (p. 615).

Principle 3: This depends on a further type of representation, *registration,* which serves as a partial surrogate for belief. An agent registers an object as being in a particular location either when the agent has encountered it in that location and not subsequently encountered it anywhere else, or when the agent successfully performs a goal-directed action on the object at that location (619). A registration is like a memory of an encountering in that it maintains the information gained in an encounter through a period in which the agent isn't directly encountering the object. If the object is moved after an agent has encountered it, the registration will not be updated, and will consequently be false. As a result, the agent will not succeed in performing a goal-directed action involving the object. In full, then, the third principle states that an agent must correctly register an object at a particular location if she is to successfully perform a goal-directed action aimed at the object. B&A say that this principle can be applied in two directions. An agent who does not correctly register an object will not be able to successfully perform actions with goals specifying that object. And if an agent does succeed in performing an action with a goal that specifies the object, it can be inferred that she has correctly registered its location (p. 617).

Principle 4: In initiating an action with a goal that specifies a particular object, an agent will approach the location at which she registers that object (p. 619). B&A claim that with this principle an infant can predict that an agent will search at the wrong location in a false belief scenario in which an object has been transferred from one location to another during her absence (p. 620).

B&A argue that a system implementing these four principles could track beliefs in a limited but useful range of circumstances[1]. An agent making use of such a system is not engaging in

[1] B&A present their account as a 'computational theory in Marr's sense', and remain neutral as to how these principles may be implemented in humans and other agents (p. 613).

mere behavior-reading (as envisioned by lean accounts), because the principles of minimal mindreading generate behavioral predictions on the basis of representations of mediating agent properties and states (field, encountering, registration). Insofar as registration is a form of stored information that has correctness conditions which allow for falsity, and is a guide to action, it has some of the properties of belief.

Nevertheless, this falls short of representing beliefs 'as such,' according to B&A. Apperly & Butterfill (2009) characterize the representation of a belief 'as such' as 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 psychological states (other beliefs, desires, emotions, preferences, etc.), and as causing and justifying actions (2009, p. 957). According to their account, 'as such' or 'full-blown' mindreading is performed by a flexible mindreading system distinct from the minimal system, with the two systems operating largely in parallel (2009, p. 964).

To understand the structure of B&A's minimal mindreading system and the reasons for postulating two systems, it is important to consider the arguments concerning efficiency and flexibility given by Apperly & Butterfill (2009; see also Apperly, 2011). 'Full-blown' mindreading, with the attributes just described, exhibits a high degree of representational power coupled with inferential holism: the mindreader can attribute any belief content that she herself is able to entertain, and belief attribution can be based on an unlimited variety and amount of information. Apperly & Butterfill (2009) claim that this flexibility is cognitively expensive – demanding executive cognitive resources – and that this kind of mindreading is hence unsuited to circumstances where processing must be fast and efficient. They include both infant mindreading and fluent everyday communication in adults as examples where

efficient mindreading is required (2009, p. 959). The minimal mindreading system is conceived in such a way as to avoid the features of 'full-blown' mindreading that make it cognitively expensive, while still achieving some significant representational ability. Thus, representational power is sacrificed by allowing only a simple fixed set of representations, and inferential holism is eliminated by relying on a simple fixed set of inferential principles. The core rationale of the theory, then, is that the two mindreading systems constitute two distinct solutions to the competing requirements of efficiency and flexibility. It is this rationale that links the Apperly & Butterfill account to other 'two systems' theories in reasoning, decision-making and social psychology (2009, p. 957; Evans & Stanovich, 2013).

The limits of the minimal mindreading system serve as the basis for identifying forms of empirical evidence that would reveal the existence of the system. Infants should be insensitive to the mode of presentation, or the way that the item is represented by the belief holder (Apperly & Butterfill 2009, p. 957; B&A, p. 621-625). This will, for instance, result in an inability to perform level-2 perspective taking, and stems from the way that minimal mindreading 'makes use of objects and their relations to agents, rather than representations of objects, to predict others' behaviours' (B&A, p. 622). Apperly & Butterfill identify the inferential holism of full-blown mindreading as a primary source of cognitive inefficiency, and suggest that infants may be unable to appreciate relations amongst multiple beliefs or between beliefs and desires (2009, p. 957). Accordingly, the more elaborated minimal mindreading scheme presented by B&A is specifically designed not to represent such relations.

As we noted in the introduction, B&A's theory provides a middle ground between lean and rich accounts of infants' precocity on non-verbal false belief tasks. A great deal of work has examined various aspects of the theory (see e.g. Scott & Baillargeon, 2009; Low & Watts, 2013; Rakoczy et al., 2014; Kovacs, forthcoming; Christensen & Michael, 2015), and our aim here is to focus on its account of goal attribution in particular. Explaining infant goal

attribution is not the main focus of B&A's theory, and, more broadly, goal attribution has not been a primary concern of research that addresses belief representation and the false-belief task. Nevertheless, goal attribution plays an important role in the false belief task, and it is also important to ensure that B&A's theory is consistent with the full range of relevant evidence.

Interpreting Principle 1

The first principle is much less clearly specified than the other three. The latter, together with their supporting explications, characterize particular forms of representation, conditions in which those representations will be invoked, and specific relations among representations. In contrast, the first principle only says that goals are represented as outcomes which are functionally-teleologically related to actions – it doesn't describe the specific form of representation employed to attribute or reason about goals. Nor is any clear account given of the conditions in which goals are attributed and the kinds of information that contribute to goal attribution, other than the proscription on information concerning psychological states. B&A only say that there is evidence that young children, non-human primates, and corvids *can* track the functions of things (2013, p. 614). But a positive theory of goal attribution must specify *how* infants identify particular outcomes as the goals of actions.

One way to view the first principle is as primarily serving a ground-clearing role. It is not intended as a positive account of goal attribution, but rather aims to suggest how infants could represent goals without representing them as dependent on intentions and other psychological states such as preferences and beliefs. The problem with this, however, is that principles 2-4 are dependent on principle 1, since principles 2-4 can only be applied in particular cases in combination with specific goal attributions. In other words, B&A's theory

of minimal mindreading is incomplete without a positive account of goal attribution, and principle 1 fails to provide such an account.

A deeper problem is that it will be difficult to develop a theoretically and empirically adequate account of goal attribution within the constraints of the approach that B&A have adopted. This is because one and the same action may have many distinct outcomes, and it is often necessary to draw flexibly upon contextual information, including information about specific agents' preferences and other psychological states, in order to identify which of the possible outcomes of an action is the agent's goal. Thus, reaching for the tap when the water is off is likely to aim at turning it on, while reaching for the tap while the water is running is likely to aim at turning it off, or perhaps at adjusting the flow. Reaching for the toy car is perhaps most likely aimed at playing with it, unless it is a tidying-up context, while reaching towards the dog may aim at patting. The challenge, then, is to explain how B&A's minimal mindreading system could take context into account in attributing goals.

In framing this challenge it will be will be useful to distinguish between *procedural* and *knowledge-based* goal attribution. This contrast follows the broad distinction between the procedural and declarative or explicit memory systems (Cohen & Squire, 1980; Squire 2004; Poldrack & Packard, 2003).[2] Knowledge-based goal attribution involves semantic knowledge and episodic memory, which are integrated and processed in working memory. We will refer to the *explicit-executive* system as the combined explicit long-term and working memory system that actively selects and processes explicit knowledge. Procedural goal attribution, in contrast, relies on procedural memory, including statistical action-effect relations, and

[2] We prefer the term 'explicit' to 'declarative', in part because it is less awkward when discussing preverbal infants. Research using deferred imitation has found explicit memory in infants as young as 6 months of age (Barr et al., 1996). In this approach the infant is shown novel actions that are demonstrated with props, and after a delay the infant is allowed to manipulate the props. The test is whether the infant re-enacts the action that has been shown. Memories evoked using this method exhibit a number of the characteristic attributes of explicit memory, including learning based on a single experience, accessibility to language (in older children), flexibility, and impaired ability when the task is given to adults with amnesia (Carver & Bauer, 2001; Bauer, 2006).

possibly specialized implicit action perception systems such as a 'motor resonance' system (e.g., Paulus et al., 2011). Since B&A's minimal mindreading system is designed to avoid flexible processing mediated by executive control, we take it that an elaborated account of goal attribution consistent with the general constraints of the theory will be largely restricted to procedural mechanisms, i.e. that it will be severely limited in the extent to which it can incorporate functional type information, since functional categories like 'scissors', 'crayon', and 'glue' are knowledge-based.

This is not to deny that semantic processing can occur outside of awareness and involuntarily. Indeed, implicit semantic priming (Neely, 1977) and implicit association tasks (Greenwald et al., 1998) provide clear evidence that it can[3]. It must be emphasized, however, that priming and implicit association tasks facilitate semantic representations in a non-task-specific way. Thus, they will inherently tend to facilitate many semantically and associatively related representations that are not relevant to the task. In contrast, the explicit-executive system can selectively activate and process task-relevant semantic information, which in the case of goal attribution would involve semantic information relevant to identifying the goal. The challenge, then, is to specify how a system like B&A's minimal mindreading system, which is designed to be largely restricted to procedural mechanisms, could achieve the requisite selectivity to ensure the activation of representations that are relevant to goal attribution in a given context.

In considering the extent to which B&A's minimal mindreading system can use contextual information for goal attribution it will also be useful to distinguish between stereotypical context information and information concerning the particularities of a given situation. In expertise research, the representation of the latter is referred to as *situation awareness* (Endsley, 1995). Situation awareness involves the construction of a situation model that captures key causal elements and relations present in the situation. Because it is

[3] We thank an anonymous reviewer for raising this point.

based on a flexible capacity for causal representation, model construction permits effective interpretation of and response to new situations. We'll call goal attribution based on integrated situation awareness *situational goal attribution*. For example, if Jenny's pencil breaks while she is writing, an onlooker with a high degree of situation awareness might predict that she will reach for a pen, since the situationally relevant causal properties of a pen are similar to those of a pencil. Note that situational goal attribution incorporates explicit knowledge and is a form of knowledge-based goal attribution. Given the proscriptions against flexibility and executive control that A&B impose, their account appears to be limited in the extent to which it can incorporate situational goal ascription

In this respect it is illuminating to consider whether B&A might appeal to Csibra and Gergely's theory of teleological action interpretation (Csibra, 2003; Gergely & Csibra, 2003). To a first approximation this seems like an appealing strategy since, in its base form, Csibra and Gergely's theory doesn't postulate the representation of mental states, but nevertheless incorporates powerful mechanisms for relating actions to goals and situations. Briefly, in this account teleological action interpretation interrelates behavior, outcome and the situation by means of a principle of efficiency, or 'rational action'. That is, an action is assumed to aim at an outcome in the most efficient way available, given the constraints of the situation. For instance, if a small ball is seen to approach a large ball via a path that seems to leap over an obstacle, it will be inferred that the goal of the small ball is to contact the large ball, and that the trajectory is an efficient means to this end, given the presence of the obstacle. If presented with a subsequent scenario in which the obstacle is not present, infants will be more surprised if the small ball follows the same 'leaping' trajectory (now over free space) than they will be by a direct path to the large ball. In its simplest form, teleological interpretation is non-mentalistic, according to Csibra and Gergely, in the sense that there is no reference to intentions, desires or beliefs but, rather, only to behavior, outcomes and situational constraints.

But although this aspect of the account is compatible with B&A's minimal mindreading system, other aspects of Csibra and Gergely's account are not consistent with B&A's theory. Specifically, Csibra and Gergely's account exhibits the kind of inferential holism that B&A's minimal mindreading system is designed to avoid. There is no in-principle restriction on the kinds of representations or inferences that can be involved in relating a behavior with the outcome and the situation, and it is claimed that novel and unusual actions and situations can be interpreted. For instance, Gergely et al. (2002) report evidence that 14-month-old infants were able to evaluate the efficiency of an action in which an adult turned on a light with her forehead, distinguishing a situation in which the adult's hands were restrained from a situation in which they were free. It is plausible that the interpretation of novel situations like this will depend on controlled semantic processing and situational goal attribution.

B&A's characterization of the minimal mindreading system also implies that information about the agent could only play a very limited role in goal attribution. The first principle, as it stands, relates a goal to an action, not to the agent performing the action. Yet the link between a taking-toys-out-of-the-toy-box-and-putting-them-on-the-floor activity and a subsequent playing-with-the-toys activity is an agent who wants to play, and who has preferences for some toys over other ones, has perceptual access to the toys, etc. By appreciating this – i.e. by identifying an agent as the organizational nexus for action – it is possible to draw on what one knows about that specific agent's prior activities, preferences and other psychological states in constraining goal attribution. We will refer to the association of a goal with an agent as *agentic linking*. In principle, B&A's account of the minimal mindreading system could be extended to explicitly include agentic linking, but it is difficult to see how such a system could make significant use of agentic linking in goal attribution. This is because it is specifically designed not to accumulate information about the activities, preferences and beliefs of specific agents to be used flexibly for goal attribution, since doing

so would require the resources of explicit memory and executive-mediated situation awareness.

In sum, B&A's minimal mindreading system is designed to avoid the representation of psychological states and the employment of flexible cognitive processes that can take into account an open-ended range of information. B&A accordingly define goals as outcomes towards which behaviors are functionally-teleologically related. The prima facie problem for this is that the same action can be performed to achieve varied goals, suggesting that it will be difficult or impossible to predictively identify goals on the basis of action type alone. Information about context can help to disambiguate goals, and B&A's theory can appeal to stereotypical context differentiation. However, it cannot appeal to mechanisms that integrate contextual information flexibly, so it will be unable to incorporate situational goal attribution or other forms of knowledge-based goal attribution. And it would therefore be severely restricted in its ability to make use of agentic linking even if it were revised to allow for the representation of preferences.

This reasoning gives us prima facie grounds to doubt that B&A will be able to construct an adequate account of goal attribution within the constraints they have adopted for the minimal mindreading system. If we recognize that infants possess explicit memory (Carver & Bauer, 2001; Bauer, 2006), and if we accept that action-goal relations often do show strong context-sensitivity, and if we further assume that the ability to interpret the actions of others is extremely important for infants, then it is reasonable to expect that they will have at least some capacity for flexible goal attribution mediated by the explicit-executive system. In the next section, we will review evidence indicating that this is indeed the case.

Evidence Against Principle 1

There is a large body of research on infant goal attribution providing strong evidence that infants relate goals to specific agents and take into account complex, idiosyncratic features of situations. In one highly influential study, for example, Woodward (1998) found that infants take prior actions into account when attributing goals. In this study, 5-month-old infants were first habituated to an event in which an agent reached for toy A in an array of two toys, A and B. In the test trial, the locations of the toys were reversed and the agent reached either for toy A at the new location, or for toy B at the original location. The main finding was that the infants looked longer when the agent reached for toy B, suggesting that they interpreted the goal of the reaching in the habituation phase as being the object rather than the location, and expected in the test trial that the reaching would have the same goal. In a more recent study based on the same paradigm, Cannon and Woodward (2012) found convergent evidence by measuring 11 month-old infants' predictive eye movements rather than looking time.

The fact that the infants in these studies expected the reach towards toy B indicates that in attributing action goals they were taking into account the previous behavior of the agent. A deflationary interpretation of this finding is that the infants simply formed an association between the agent and toy A which was broken in the test trial. An alternative interpretation is that the infants saw the selection of toy A in the habituation phase as a *contrastive choice* that revealed a preference for toy A in comparison with toy B, and they expected this preference to guide the agent's actions in the test situation.

A study by Luo and Baillargeon (2005) supports the latter interpretation. 5-month-old infants were habituated to a self-propelled box approaching a target (a cone). In a non-contrastive condition there was only one possible target during the familiarization phase, while in a contrastive choice condition there was a second possible target (a cylinder), which

the box never approached. In the test phase a cone and cylinder were present. If the infants experienced the contrastive choice condition during the familiarization phase they expected the box to again approach the cone. But if they experienced the non-contrastive condition they had no expectation concerning which object the box would approach. This undermines a simple association interpretation because there is as much reason to expect an association between the agent and the cone to form in the non-contrastive condition as in the contrastive choice condition.

A later study examined whether 12.5-month-old infants are sensitive to the agent's perceptual access when there is apparent contrastive choice. Luo and Baillargeon (2007) employed a habituation phase that included a visible object condition in which the agent could see that there was a second object. In a hidden object condition a second object was present but the agent couldn't see it. In the test trial the agent selected one of the two objects. If the infants had experienced the visible object condition they expected the agent to maintain the same goal, whereas if they had experienced the hidden object condition they had no expectation. In a more recent study with 6-month-olds, Kim & Song (2015) reported convergent evidence using predictive eye movements rather than looking time as a measure. Luo (2011) showed further that infants not only take into account perceptual access in detecting contrastive choice, they can also take into account the beliefs of the agent. She found that 10-month-olds did not attribute a preference for an object to an agent if the agent had been interacting with the object but believed (truly or falsely) that no other objects were present.

Other studies have examined infant sensitivity to higher order goal structure, in particular whether they interpret an initial action in a multi-action sequence as being aimed at the overall outcome. Sommerville & Woodward (2005) employed a task that involved pulling a piece of cloth to obtain a toy sitting on it that is out of reach. The agent faced two pieces of cloth of different colors, and on each cloth there was a toy out of reach, different from the toy

on the other cloth. In habituation trials the agent pulled one of the cloths towards her and grasped the toy. The question at issue was whether infants interpreted the action of pulling the cloth as having the goal of obtaining the particular object on it. In test trials the location of the toys was reversed. The agent either grasped the same cloth as previously, or grasped the other cloth, which had the same toy as was previously attained. In neither case was a toy touched. Twelve-month-old infants were more surprised when the agent grasped the same cloth, indicating that they interpreted the toy as being the target of the action. This result also suggests that infants represent causal relations by which actions achieve outcomes, including mediative relations in which an agent acts on an object without physically touching it. In a variation of the experiment the causal relation was broken: the toys were beside rather than on the cloths. In habituation trials the agent first pulled the cloth then reached for the toy beside the cloth. In test trials toys were swapped and the agent either reached for the same or the other cloth. Infants in this condition did not respond with longer looking times when the same cloth (with the new toy adjacent) was grasped.

A study with 13.5-month-olds by Song, Baillargeon & Fisher (2005) also indicates an understanding of hierarchical action relations based on causal understanding. They first presented infants with three familiarization trials in which an agent grasped an object on the floor of an apparatus and slid it back and forth. Various objects (a toy fish, a box, and a shoe) were used. The infants were then shown a display with two identical toy trucks resting next to each other on the apparatus floor. The truck on the right was in a short frame, making it impossible for the agent to slide the truck back and forth. The truck on the left was in a longer frame that had enough space for the agent to slide the truck back and forth. Finally, on the test trial, the agent grasped one of the trucks – either the one in the long frame or the one in the short frame. The main result was that infants who saw the agent grasp the truck in the short frame looked reliably longer. This indicates that they attributed to the agent the goal of sliding

an object back and forth, and understood that this goal could only be achieved with the truck in the long frame.

These studies suggest strongly that infants in the second year of life attribute goals to agents rather than to actions, and do so in a manner that is constrained by the preferences and epistemic states of specific agents (i.e. making use of agentic linking). However, they don't rule out the possibility that infants take contrastive choice to reveal the value of the object rather than a preference of the agent. Adults often interpret contrastive choice both as revealing a preference of the chooser and as indicating that the preferred object is (or might be) valuable. It's conceivable, however, that infants might assign goals to actions and values to objects, but not assign goals and preferences to agents. To eliminate this possibility, Buresh and Woodward (2007) employed a different agent in the habituation and test phases of an experiment using the contrastive choice design. If goals and preferences are attributed to agents, then a contrastive choice by agent A when presented with two objects should not influence the infant's expectations for the subsequent choice of agent B when presented with the same two objects. If infants interpret contrastive choice as indicating the value of an object, and assume that any agent will select an object thus shown to be valuable, then they should expect agent B to select the same object as agent A. Buresh and Woodward found that 9 and 12 month old infants had no expectations for the choice of agent B, indicating that they interpreted contrastive choice as revealing a preference of the specific agent performing the contrastive choice.

Finally, the findings from a study by Spaepen & Spelke (2007) indicate that 12 month-olds draw upon generic knowledge of categories of objects in attributing goals to agents. Specifically, when an agent had preferentially chosen a red female doll over a blue truck during familiarization, and was then faced with the choice between a blue male doll and a red truck in the test phase, the infants expected her to have the goal of grasping the doll. This reveals a capacity for active selection of contextually relevant semantic information

(knowledge-based goal attribution), which is difficult to account for within the constraints which B&A's theory imposes on the minimal mindreading system.

Taken as a whole, this body of research supports the view that goal attribution in infancy is informed by generic semantic knowledge (knowledge-based goal attribution) and is sensitive to the specific causal structure of the situation (situational goal attribution). Moreover, infants also draw upon information about agent-specific preferences in attributing goals (i.e. making use of agentic linking). As we argued in the previous section, these abilities are difficult to explain within the constraints adopted by B&A in their account of the minimal mindreading system. Furthermore, the contrastive choice experiments indicate that infant goal attribution is sensitive to the epistemic situation of the agent, taking into account both perceptual access and beliefs. A primary motivation for B&A's teleological account of goal attribution is to avoid the need to make the representation of goals dependent on other psychological states, like desires and beliefs, yet this evidence suggests that infants do treat goals, preferences and beliefs as interrelated and mutually influencing.

Problems in Applying Principles 2-4

The problems with principle 1 have consequences for the rest of B&A's account because principles 2-4 depend on appropriate goal attributions. We can illustrate this by considering some specific cases. The false belief task devised by Träuble et al. (2010) involves a complex, novel situation in which correct belief attribution depends on situationally sensitive goal attribution. In this study, Träuble and colleagues showed that infants could correctly ascribe true and false beliefs to an agent about the location of a ball when the agent manipulated an apparatus without visual access. Specifically, the apparatus was 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. In a true belief condition the agent

manipulated the beam herself while facing forward and able to see the transfer of the ball. In a false belief condition the agent's back was turned and the beam was manipulated without her input, resulting in the transfer of the ball. In a manipulation condition the agent again faced away from the apparatus but manipulated the beam herself, causing the ball to transfer between boxes. In each condition two different outcomes were contrasted: either the agent reached for the ball in the original box (wrong location) or in the new box into which the ball had rolled (correct location). The key result was that 15-month-olds expected the agent to reach for the correct location in the true belief and manipulation conditions, but not in the false belief condition.

One possible interpretation of these results, which is consistent with the findings discussed in the previous section, is that the infants attributed to the agent a causal understanding of the balance beam and used her manipulation of the beam as a basis for ascribing a belief about the location of the ball to her. This interpretation is not compatible with B&A's account: it recognizes physical manipulation as a source of beliefs, and it also recognizes interactions between beliefs: the belief about the ball's location is mediated by a belief about how the apparatus operates. B&A offer a different interpretation. They say (2013, footnote 12, p. 617) that this case can be covered by the clause in principle 3 which stipulates that it can be inferred that the agent has registered the location of an object if that agent has successfully performed a goal-directed action involving the object. In other words, infants don't represent physical manipulation as a source of beliefs, or relations between beliefs about the apparatus and beliefs about the ball, but they do infer beliefs on the basis of successful manipulation. In effect, because the agent performed a successful action which had the ball as its target, the agent must *somehow* have registered the location of the ball.

For B&A's explanation to work, though, the infant must identify the goal of the agent's action as moving the ball to the new box. Notably, the agent doesn't directly act on the ball or interact with the ball at the new location until she reaches into the second box. And

in manipulating the beam with her back turned the agent might have had other goals. She might have simply wanted to raise the beam, for example. How could an infant employing principle 1 specifically pick out moving the ball to the new box as the goal of the action?

The action is novel and unusual, so the goal can't be identified as the procedural outcome of an established action type. One possibility is that the infant classifies the action as a type during the familiarization phase, where the action was repeatedly demonstrated. In these trials, the agent's gaze did conspicuously follow the movement of the ball, possibly marking for the infant the movement of the ball as the goal of lifting the beam. The action was only demonstrated four times, however. Implicit learning characteristically occurs through lengthy exposure (Eichenbaum, 2003; Squire, 2004). It is thus not clear how B&A might explain this result, and the alternative interpretation – that the infants identify the goal by attributing to the agent a causal understanding of the apparatus – is strengthened by the results described in the previous section, which indicated that infants are sensitive to the causal structure of action.

The problems with principle 1 also result in difficulties explaining standard change-of-location false belief tasks. In the conventional scenario, an agent (sometimes called Sally) places an object (such as a doll) in a box and then leaves the room. While Sally is gone, another agent (Anne) moves the doll to a second box. Sally subsequently returns and approaches one of the two boxes. The problem for B&A's theory is that when Sally re-enters the room the infant must ascribe to her not only a belief about where the doll is, but also the goal of obtaining the doll. More specifically, the fourth principle says that in initiating an action with a goal that specifies a particular object, an agent will act as if the object is at the location where she has registered it. This is intended to explain why the infant expects Sally to approach the box where she falsely believes the doll to be. But to generate this prediction the doll must be specified as the goal of Sally's action, and there is no clear basis for this ascription in the first principle as it is currently formulated. If goal attribution is based on

procedural action type, and procedural action type is determined by movement pattern, then the infant should simply ascribe to Sally the goal of engaging with whichever box she approaches. Nor is there any stereotypical structure in the immediate context as Sally re-enters the room that links her action to the goal of obtaining the doll, in the way that a running water tap can indicate that the reaching aims to turn it off.

These considerations don't show definitively that B&A's approach cannot work: the characterization of the minimal mindreading system can be extended, and it is an open question what any such extensions might or might not be able to explain. But the account is designed to avoid appeal to flexible explicit-executive cognitive processes and the representation of interdependencies amongst psychological states – two key features that B&A associate with 'full-blown' mindreading. Especially when taken as a whole, the results that we've surveyed reveal a high degree of flexibility in infant goal attribution. We suggest that it will not be easy to develop a parsimonious explanation of these results which does not involve explicit-executive cognitive processing.

This point can be reinforced by considering some additional false belief experiments that appear to involve agentic linking and the use of semantic knowledge. A paradigm employed by Surian et al. (2007) uses contrastive choice to establish that one of two objects is the goal of the agent: the agent sees both an apple and a piece of cheese being placed behind screens, and consistently approaches the cheese. Here, infants appear to take into account agent preferences in order to determine what the agent's goal is in the false belief condition. Other studies seem to show flexible integration of semantic knowledge. For example[4], Scott et al. (2010) presented 18-month-olds with a task in which one agent watched while a second agent demonstrated that a target object rattled when shaken. In the test phase, the first agent had the opportunity to choose between two objects to produce the rattling sound. One of the objects was similar in appearance to the target object while the other was dissimilar. The

[4] We thank an anonymous reviewer for pointing out the relevance of this study.

infants expected the second agent to select the object that was similar, even though they knew that in fact it was the dissimilar object that rattled. This suggests that the infants were employing semantic knowledge to form the expectation that the agent would use similar appearance to the model object as a guide to which test object would rattle.

Two Kinds of Response to These Problems

To review, the basic problem is that principle 1 of B&A's minimal mindreading system does not provide a positive account of goal attribution. Principles 2-4 depend on the attribution of specific goals, and without a positive account of goal attribution the theory fails to account for infant expectations in the false belief tasks it is intended to explain. In section 3 we argued that the overarching assumptions that B&A have adopted require that goal attribution by the minimal mindreading system not include flexible, controlled knowledge-based processes. This, in turn, rules out situational goal attribution and severely limits the potential to make use of agentic linking to inform goal attribution. The contrastive choice experiments reviewed in section 4 provide a substantial body of evidence indicating that infants do indeed make use of agentic linking and situational goal attribution. And in section 5 we argued that procedural goal attribution will struggle to explain goal attribution in the balance beam task and standard change-of-location false belief tasks.

There are two main ways that B&A might respond to these difficulties. The first is to elaborate the first principle along the same lines as the second, third and fourth principles. That is, B&A might postulate a specialized representational system for goal attribution. This would need to explain how goal attribution works for novel actions based on initial acquaintance or limited exposure, and explain the tracking of goals across extended action sequences. For example, it might be specified that, on the first performance of a novel action, goal identification is based on movement patterns and/or eye gaze. On subsequent occasions

the same goal will be ascribed unless there are cues indicating that the context is different, in which case there will be no goal attribution. In certain conditions, action directed at an object will be taken to establish an enduring preference for the object which influences subsequent actions.

The threat facing this approach is that specifying in detail the conditions in which goals are and are not attributed will require a complex and ad hoc set of representations and representational relations. Appropriately tracking goals across multiple actions and through mediated causal relations in particular presents a difficult challenge. As we have seen, infant goal attribution appears to be sensitive to causal relations – infants attribute a goal to an action when the causal relation to the outcome is intact, and do not attribute the goal to the action when the causal relation is broken. But by assumption the minimal mindreading system lacks the flexibility and control required for integrated situation awareness, and it consequently does not have access to causal information concerning action-outcome relations. To capture this sensitivity, the goal attribution system would require proxy cues that correspond to intact and broken causal relations. But since such relations can be extremely diverse, it is hard to see what such cues might be. For this reason we doubt that this approach can succeed.

The second type of response that B&A might adopt is to abandon the idea that infant mindreading is strongly encapsulated. Instead, mindreading might involve an interplay in which a variety of specialized systems are integrated via executive cognition. Thus, specialized systems for causal representation, agent tracking, the representation of agent's attitudes, and action structure might all take input from and contribute to generalized situation awareness. The overall integration of information might depend not on 'hard coded' principles but on flexible cognitive inferences facilitated by learning. This approach would constitute a fundamental change in orientation, however. It would require abandoning the idea that the disparity between infant false belief performance using implicit and explicit measures is because the former depends on a separate mindreading system that distinctively does not

employ flexible, explicit cognitive processes. It would also require reconsideration of the core reasoning concerning efficiency and flexibility on which B&A base their two systems account.

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., 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

Barr, R., Dowden, A., & Hayne, H. (1996). Developmental changes in deferred imitation by 6- to 24-month-old infants. *Infant Behavior and Development*, *19*(2), 159–170. http://doi.org/10.1016/S0163-6383(96)90015-6

Bauer, P. J. (2006). Constructing a past in infancy: a neuro-developmental account. *Trends in Cognitive Sciences*, *10*(4), 175–181. http://doi.org/10.1016/j.tics.2006.02.009

Buresh, J. S., & Woodward, A. L. (2007). Infants track action goals within and across agents. *Cognition*, *104*(2), 287–314. doi:10.1016/j.cognition.2006.07.001

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

Cannon, E. N., & Woodward, A. L. (2012). Infants generate goal‐based action predictions. *Developmental Science*, *15*(2), 292-298.

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

Carver, L. J., & Bauer, P. J. (2001). The Dawning of a Past: The Emergence of Long-Term Explicit Memory in Infancy. *Journal of Experimental Psychology: General*, *130*(4), 726–745.

Christensen, W. and Michael, J. (in press) From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science*, *210*(4466), 207–210. http://doi.org/10.1126/science.7414331

Csibra, G. (2003). Teleological and Referential Understanding of Action in Infancy. *Philosophical Transactions: Biological Sciences*, *358*(1431), 447–458.

Eichenbaum, H. (2003). Learning and memory: brain systems. In L. R. Squire, S. K. McConnell, J. L. Roberts, N. C. Spitzer, & M. J. Zigmond (Eds.), *Fundamental Neuroscience* (pp. 1299–1248). Elsevier.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64. doi:10.1518/001872095779049543

Evans, Jonathan St BT, & Stanovich, K. (2013) Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science* 8(3): 223-241.

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

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292. http://doi.org/10.1016/S1364-6613(03)00128-1

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Heyes, C. (2014). False belief in infancy: a fresh look. *Developmental Science*, n/a–n/a. doi:10.1111/desc.12148

Kim, E. Y., & Song, H. J. (2015). Six-month-olds actively predict others' goal-directed actions. *Cognitive Development*, *33*, 1-13.

Kovács, A. (in press). Belief-files in theory of mind reasoning. *Review of Philosophy and Psychology*.

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

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science*, 24, 305–311. doi:10.1177/0956797612451469

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition*, 121(3), 289–298.

Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old-infants. *Psychological Science*, 16, 601–608.

Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, 105(3), 489–512.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226.

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

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

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia*, *41*(3), 245–251. http://doi.org/10.1016/S0028-3932(02)00157-4

Rakoczy, H., Fizke, E., Bergfeld, D., & Schwarz, I. (2014). Explicit theory of mind is even more unified than previously assumed: belief ascription and understanding aspectuality emerge together in development. *Child Development*.

Rubio-Fernández, P., & Geurts, B. (2012). How to pass the false-belief task before your fourth birthday. *Psychological Science*, 0956797612447819.

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

Scott, R. and Baillargeon, R. 2009: Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, 80, 1172-1196.

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

Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition*, *95*(1), 1–30. doi:10.1016/j.cognition.2003.12.004

Song, H., Baillargeon, R., & Fisher, C. (2005). Can infants attribute to an agent a disposition to perform a particular action? *Cognition*, *98*(2), B45–B55. doi:10.1016/j.cognition.2005.04.004

Southgate, V., Senju, A., & Csibra, G. (2007). Action Anticipation Through Ascription 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

Spaepen, E., & Spelke, E. (2007). Will any doll do? 12-month-olds' reasoning about goal objects. *Cognitive psychology*, *54*(2), 133-154.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. http://doi.org/10.1037/0033-295X.99.2.195

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177. doi:10.1016/j.nlm.2004.06.005

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

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