Action Understanding in Infancy: Do Infant Interpreters Attribute Enduring Mental States or Track Relational Properties of Transient Bouts of Behavior?

Marco Fenici,[a] Tadeusz W. Zawidzki[b]

[a] Department of Literature and Philosophy, University of Florence
[b] Department of Philosophy, George Washington University

We address recent interpretations of infant performance on spontaneous false belief tasks. According to most views, these experiments show that human infants attribute mental states from a very young age. Focusing on one of the most clearly worked out, minimalist versions of this idea, Butterfill and Apperly's (2013) "minimal theory of mind" framework, we defend an alternative characterization: the minimal theory of rational agency. On this view, rather than conceiving of social situations in terms of states of an enduring mental substance animating agents, infant interpreters parse observed bouts of behavior and their contexts into goals, rational means to those goals, and available information. In other words, the social ontology of infant interpreters consists in goal-directed, (mis- or un-) informed bouts of behavior, by non-enduring agents, rather than agents animated by states of enduring, unobservable minds. We discuss a number of experiments that support this interpretation of infant socio-cognitive competence.

Keywords: mindreading, social cognition, action prediction, false belief, cognitive development

1. Introduction

In the wake of a flurry of recent experimental results in developmental psychology, there is currently a vibrant debate about how to characterize the so-

Corresponding author's address: Tadeusz W. Zawidzki, Philosophy, Phillips Hall 524, George Washington University, 801 22nd St. NW, Washington DC 20052, USA. Email: zawidzki@gwu.edu.

Studia Philosophica Estonica (2016) **9.1**, 237–257
Published online: December 2016
Online ISSN: 1736–5899
www.spe.ut.ee
http://dx.doi.org/10.12697/spe.2016.9.1.10

cial cognition of pre-verbal infants in the second year of life.[1] Since the earliest research on *elicited* false belief tasks (Baron-Cohen et al. 1985; Wimmer and Perner 1983), it had been thought that children do not attribute representational mental states to others until age four (Wellman et al. 2001). These sorts of experiments can be run only on verbal children, since they involve *asking* subjects about agents' mental states or future behaviors. However, primarily employing a violation-of-expectations looking-time paradigm, more recent experiments have shown that infants *spontaneously* look longer, and hence show greater surprise, when an uninformed or misinformed agent acts in an informed manner than when she acts in an uninformed or misinformed manner, and when an informed agent acts in an uninformed or misinformed manner than when she acts in an informed manner (Onishi and Baillargeon 2005; see Baillargeon et al. 2010, for a review).

This very robust pattern of apparently inconsistent results has provoked a spectrum of theoretical responses. At one extreme, many of the pioneers of the spontaneous response paradigm argue that it shows that, by early in the second year of life, human infants are deploying the *same* concepts of mental states, including preferences and false beliefs, to interpret behavior, as older children who pass the elicited response tasks after age four (Luo and Baillargeon 2010). The lag in passing elicited response versions of the false belief task is attributed to immature domain-general capacities, like working memory, attention, and response inhibition (Scott and Baillargeon 2009), or to immature linguistic capacities affecting children's interpretations of verbal requests for behavioral predictions or mental state attributions (Carruthers 2013; Helming et al. 2016). At the other extreme, some theorists argue that infant behavior on spontaneous response false belief tasks is absolutely *no* evidence for an early developing, specifically socio-cognitive capacity. Infant responses can be explained entirely in terms of associations between agents and locations or objects (Ruffman 2014; Ruffman and Perner 2005). Finally, some theorists argue for a compromise position: infant behavior in spontaneous response false belief tasks is evidence for a dedicated socio-cognitive capacity, but not the *same* socio-cognitive capacity as four-year-old children deploy when they pass elicited response false belief tasks. Spontaneous response false belief tasks trigger a "minimal theory of mind" which, although it involves the deployment of a type of mental state concept, does not involve the deployment of fully meta-representational mental state concepts,

[1] Marco Fenici wrote section 1; Tad Zawidzki wrote section 2. Sections 3, 4, and the Conclusions were written by both of the authors together.

of the kind deployed by four-year-olds in elicited response false belief tasks (Apperly and Butterfill 2009; Butterfill and Apperly 2013).[2]

Close attention to, and careful reflection on the infant data are essential to choosing between these different proposals. On the one hand, no one assumes that infants must be able to display belief-tracking capacities in *every* interpretive context. Thus, their apparent absence in *some* experimental situations does not directly settle whether this is caused by competence as opposed to performance limitations (Bloom and German 2000; Fodor 1992). On the other hand, because intentional action manifests itself in observable physical behavior, it is always possible to interpret apparent manifestations of infant belief-tracking capacities as evidence for either mentalist or behaviorist interpretive capacities (Buckner 2014; Hutto 2015; Povinelli and Vonk 2004). Resolving this debate thus requires defining a precise criterion to indicate what socio-cognitive capacities infants display in particular experimental contexts (Fenici 2015, 390–392).

Below, we focus on Butterfill and Apperly's (2013) "minimal theory of mind" proposal because of its admirable clarity in identifying empirically determinable "signature limits" meant to distinguish infant mindreading from full-blown, meta-representational theory of mind. According to Butterfill and Apperly, rather than attributing fully representational beliefs and preferences to agents, infant interpreters attribute goals, encounterings, and registrations. Roughly, the goals of behaviors displayed in infant experiments are fully public states of affairs that agent behavior brings about. Encounterings are transparent relations between agents and worldly states, like lines of sight on objects at locations. Registrations derive from encounterings: an agent who has encountered object O at L, will register this information, and continue to be guided by it in the absence of a new, incompatible encountering, e.g., line of sight on O at L′. Since registrations can be *insensitive* to changes in situations, e.g., when something happens while the agent is not looking, agents guided by registrations can behave in ways that make sense only relative to non-actual situations. Thus, the attribution of registrations enables infant interpreters to exhibit behavior that is typically

[2] Heyes (2014) argues for a fourth possibility: that infants' selective attention in spontaneous response false belief tasks can be explained by the operation of domain general processes, and depends on the novelty of low-level features of the scenes presented to infant subjects, such as colors, shapes, and movements. We similarly believe that domain general learning may underlie infants' progressively refined sensitivity to other agents' goal-directed behavior (see Fenici 2014). In contrast to Heyes (2014), however, we propose that infants' learning progressively comes to focus on the domain of specifically goal-directed behavior. Thus, unlike Heyes' view, our view does not depend on dismissing this whole experimental paradigm, on the basis of alleged experimental confounds.

diagnostic of false belief attribution, without attributing full-blown false beliefs.

One of the key signature limits that Butterfill and Apperly identify for social cognition guided by minimal theory of mind involves opacity: they conjecture that infants employing only minimal theory of mind will *not* be able to track differences in the modes of presentation with which different agents track the same objects. Since the minimal mindreading analog of false belief, i.e., registration, is a transparent relation between an agent and an object at a location, infants should show no sensitivity to differences in the appearances by which different agents track the same objects.

Another signature limit involves holism: there should be limits to the degree of inferential "promiscuity" that infants employing minimal theory of mind associate with the states they attribute. Full-blown mental states can, in principle, interact with whole systems of other mental states, as well as a wide variety of inputs (different sense modalities, verbal interactions), yielding different behaviors as a result of different interactions. There seems to be no limit to such potential inferential promiscuity, especially with some categories of mental states, like the propositional attitudes. But, according to Butterfill and Apperly, this is not the case for the states attributed by minimal mind readers, like "registrations" and "goals." Although registrations and goals can interact with some flexibility, i.e., the same registrations will yield different behaviors depending on goals, and vice versa, there are important limits to this.

Butterfill and Apperly's signature limits to minimal theory of mind thus imply specific predictions about how infants will perform on spontaneous response tasks. There are already experimental paradigms exploring this question. Specifically, there is evidence that infant social cognition is bound by Butterfill and Apperly's first signature limit, a failure to appreciate opacity (Low and Watts 2013). In contrast, there seems to be strong evidence that infant interpreters are *not* bound by the second signature limit: they seem surprisingly flexible in the degree of inferential promiscuity they associate with the mental states they attribute. For example, they seem capable of linking specific beliefs with an indefinite variety of perceptions originated through different modalities, including sight (Onishi and Baillargeon 2005; Southgate et al. 2007), touch (Träuble et al. 2010), and verbalized information (Song et al. 2008).

While these studies address some issues relevant to the interpretation of infant social cognition, we believe that the evidence they provide is not enough to show that preverbal infants deploy *any kind* of mental state concept when they pass spontaneous response versions of false belief tasks. In particular, there is a feature of mature mental state concepts that has received very little attention in the experimental literature, yet which there are good empirical reasons to doubt infants appreciate. This is the fact that mature concepts of mental states are concepts of *states*. At a minimum, a state must be one of a number of alternative possible states into which some enduring object can enter. Presumably, the relevant object for mental states is an agent's *mind*, where this is understood as an unobservable, enduring, causal nexus, responsible for observable behavior.

In what follows, we argue that none of the evidence from spontaneous response false belief tasks with preverbal infants shows that infants conceive of agents as animated by enduring, unobservable minds, of which the beliefs and preferences they allegedly attribute are supposed to be states. Furthermore, we argue that there is evidence that they do *not* conceive of agents in this way, i.e., as animated by minds which can enter different kinds of causally potent states, like beliefs and preferences.

How else might infants conceive their social world if not in terms of an ontology of agents animated by enduring minds that enter different kinds of mental states? Here is another possibility. Perhaps infant interpreters employ an ontology of disjoint, temporally limited and isolated bouts of behavior. These bouts of behavior can have different goals, and can be informed, uninformed, or misinformed. But, we want to suggest, there is no evidence that infants represent them as causal consequences of states of an unobservable, enduring mental substance, animating individual agents. Infants simply parse observable bouts of behavior and their spatiotemporally limited contexts into obvious goals, rational means to those goals, and available information, predicting their future trajectories based on these attributed relational properties. This should be enough to explain their capacity to distinguish between the future courses of informed, misinformed, and uninformed behaviors, without assuming that they deploy mature concepts of mental states.

Below, we proceed as follows. In Section 2, we provide more detailed characterizations of the central concepts we claim infants deploy, especially the concept of a goal-directed bout of behavior that can be informed, uninformed, or misinformed. In Section 3 we review positive, existing evidence that infant social cognition, as measured via spontaneous response tasks, shows precisely the signature limits one would expect, if they operate with an ontology of disjoint, spatiotemporally limited, (un- or mis-) informed, goal-directed bouts of behavior, rather than an ontology of agents animated by states of enduring mental substances. Section 4 discusses the question of whether there is a way of experimentally resolving this disagreement about the social ontology of preverbal infants.

2. The goal directedness and informedness of bouts of behavior

Most discussion of social cognition makes a tacit assumption: behavioral anticipation is based either on the attribution of mental states, or on extrapolation from concrete, observable behaviors. Since the latter disjunct is highly implausible, given the variety of observable behaviors that appear to yield similar predictions, and the variety of predictions that appear to draw on similar observable behaviors, most conclude that the social cognition of human adults and infants, as well as of many nonhuman species, must be mentalistic. But this, we submit, is based on a false dilemma: there are options other than behaviorism and mentalism in interpreting socio-cognitive capacities, specifically, those revealed in the infant data. In this section, we clarify what we mean when we say infants conceptualize bouts of behavior as goal directed and (mis- or un-) informed, yet not as caused by beliefs and preferences.

Consider first what it is for a behavior B to have goal G. Of course, this cannot be equivalent to some concrete, observable, intrinsic property of B, since an indefinite variety of behaviors can all have goal G. However, it does not follow that it must be equivalent to possessing a mentalistic etiology, i.e., roughly, being caused by an unobservable representation of G. After all, *having* a goal is not equivalent to *representing* that one has a goal.

Here is an alternative: to say that B has G as its goal is to say that, of all relevant, observable alternatives in this context, B is the most efficient means of bringing about G. This way of understanding what it is for a behavior to have a goal is abstract without being mentalistic—as also claimed by Gergely and Csibra, the first advocates of this form of "teleological stance" (Csibra et al. 1999; Csibra et al. 2003; Gergely and Csibra 2003). An indefinite variety of behaviors that differ in their intrinsic, observable properties, can all share the relational property of having G as their goal, in this sense, because, in different contexts, different behaviors will count as the most efficient means of bringing about G.

Significantly, infants can predict the future course of observed behavior by interpreting it in this non-mentalistic way. It would be sufficient for them to generate candidate observable goals for an observed behavior, and select as *the* goal of the behavior the one relative to which the behavior appears to be the most efficient means. For instance, infants may evaluate the most efficient means of accomplishing various goals by detecting the statistical frequency of different observed patterns of behavior as well as by exploiting the background knowledge they have acquired from experience of acting directly (Fenici 2014; Fenici and Carpendale submitted; Ruffman 2014); or they may simply run their own planning systems off line, and use the generated behavioral predictions to anticipate interpretive targets' future behaviors (Nichols and Stich 2003). This is not equivalent to attributing representations of goals, or simulating interpretive targets' mental states. Infant interpreters need not think of themselves as accessing their own planning system or background information, and then projecting representations of goals they generate onto the minds of interpretive targets.

Consider now how one can represent a bout of behavior as informed, or uninformed, or misinformed, without conceiving of it as caused by a (mis-) representation of some situation. This appears to be more challenging than characterizing goal-directedness non-mentalistically. Indeed, infants' apparent sensitivity to the fact that behaviors can be misinformed is the reason why many are inclined towards a mentalist interpretation of the data. In fact, even advocates of minimalist interpretations, like Butterfill and Apperly, assume that when an infant interpreter predicts that an agent A will *not* take into account the fact that an object O has been moved from L to L′ without its knowing, she must be attributing an intervening variable as the cause of this misinformed behavior. This assumption appears justified, since the state to which the infant appeals in predicting A's misinformed behavior, i.e., O's presence at L, no longer obtains; so, *that* state cannot be what the infant has in mind when predicting the target's current behavior. Rather, the infant must be attributing to A the registration that O is at L, an unobservable, intervening variable that causes A's behavior. This qualifies the infant's interpretation as mentalistic, and is the reason why Butterfill and Apperly call their minimalist interpretation of infant competence, a minimal theory of *mind*.

It is unclear, however, why the case of misinformed behavior should be any different from the case of any other kind of goal-directed behavior. All goal-directed behavior is predicted on the basis of a non-actual state: goal G does not obtain yet; so, G cannot help predict A's behavior. Therefore, the reason for thinking that misinformed behavior requires the attribution of an intervening, mentalistic cause, i.e., the interpretive target's behavior is sensitive to a non-actual situation, equally supports the claim that *any* goal-directed behavior requires the attribution of an intervening, mentalistic cause. But the latter seems neither plausible nor necessary. It is *prima facie* implausible, given how early in development infants seem sensitive to the goals of behaviors, and how widespread such sensitivity is among non-human species. More significantly, it is also unnecessary: given our earlier, non-mentalistic characterization of what it is for behavior B to have G as its goal, it is clear that this can be conceptualized without the attribution of an intervening, mentalistic cause. Similarly, neither Butterfill and Apperly (2013), nor Gergely and Csibra (2003) have ever claimed that *all* sensitivity to goal-directed behavior requires the attribution of such intervening variables. Therefore, just as the sensitivity to non-actual states of goal-directed behavior need not imply that successful prediction of such behavior requires the attribution of intervening states, it is possible that the sensitivity to non-actual states of *misinformed* behavior likewise need not imply this.

Here is our non-mentalistic explanation of the infant capacity to predict misinformed behavior. Suppose infants begin with the teleological stance: they expect agent behaviors to constitute the most efficient means to goals among alternatives available in context. These expectations will often be confuted, due to mismatches between the background assumptions of infant interpreters and agents they interpret. Sometimes agents will have non-obvious goals, which infant interpreters will not even consider as possible candidates to be weeded out using the "efficient means" heuristic. Sometimes agents will have different background knowledge about how to accomplish goals, and hence make different assumptions about which behavior is the most efficient means. Significantly, sometimes agents will also have access to different information about the layout of objects in the environment, with implications for which behaviors are the most efficient means to some goal.

Given the drive to minimize prediction errors, infant interpreters will notice behavioral and situational invariants that accompany such errors. For example, they might notice early on that interpretive targets with no line of sight on the displacement of an object relevant to some goal will subsequently fail to pick the most efficient behavioral means to that goal. Furthermore, they might notice that the behavior such an interpretive target picks remains the most efficient means to the goal relative to the object position on which the target last had a line of sight, prior to displacement.

If this captures infant socio-cognitive development, it makes possible non-mentalistic characterizations of informed, uninformed, and misinformed behaviors. Suppose that infant interpreters assume the following. Behavior B, with goal G, is informed by situation S just in case B constitutes the most efficient means to G only relative to S. To know by which S B is informed, infants must simply monitor the epistemically relevant environmental relations of whichever agent, A, performs B. For example, suppose B, performed by A, is the most efficient means to G only relative to the situation S, in which object O is at L. If O remains at L throughout, and A has a constant, direct line of sight on O, the infant interpreter will predict that B will be informed by S, and successfully achieve G. In contrast, if O is moved from L, yielding a new situation, S', and A does *not* have a constant direct line of sight on O, the infant interpreter might predict that B will be uninformed, i.e., insensitive to S', and hence, fail to accomplish G.

Or, perhaps the infant interpreter is more sophisticated. Perhaps she will assume that B is *misinformed* by situation S, i.e., O at L, in which A last had a line of sight on O. In that case, the infant will predict that B will constitute the most efficient means to G relative to (the now anachronistic) S. This is the sort of case that is typically used to diagnose the presence of the false belief concept. However, we see here that such cases need not imply an understanding of false belief, understood as a mental state, of some unobservable causal nexus, responsible for the behavior. Instead, infant interpreters might just tag a behavior as misinformed based on the fact that its agent lacked epistemic access to recent alterations relevant to the behavior's goal.

To sum up: nothing in the available empirical data demonstrates that infants' interpretations of intentional action require them to attribute *representations* of goals, as opposed to generating candidate observable goals for an observed behavior, and selecting as *the* goal of the behavior the one relative to which the behavior appears to be the most efficient means—that is, applying the teleological stance to the observed behavior. Similarly, nothing in the empirical data compels the conclusion that infants attribute intervening mental states in spontaneous response false belief tasks. Rather, infants may simply notice that agents with no line of sight on, or other observable epistemic access to the displacement of an object relevant to some goal, will subsequently select the most efficient means to the goal *relative to the object position of which they were last informed*, prior to displacement.

3. Assessing the empirical evidence

Above, we have introduced what we may call a minimal theory of rational agency, according to which infants form expectations about others' behavior by tracking the goals of disjoint, temporally limited and isolated bouts of behavior, and also paying attention to short-lived behavioral and situational invariants affecting whether or not such behaviors are appropriately informed. We now discuss how we might experimentally establish whether or not an infant interpreter is operating with such concepts of goal-directed, (mis- or un-) informed behaviors, rather than with concepts of enduring minds that enter states of belief and preference.

We first note that even some of the most persuasive experimental evidence in favor of preverbal infant mastery of mental state concepts is equally compatible with our view that infant interpreters conceptualize their social worlds only in terms of goal directed and (un- or mis-) informed bouts of behavior. Consider, for example, Scott and Baillargeon's (2009) experiment with two toy penguins. One penguin comes in two pieces, but can be assembled to look like the other, one-piece penguin. The goal of the agent infants interpret is to place a key in the two-piece penguin. The agent comes upon the following scene, which the infant has just observed being set up: the two-piece penguin assembled to look like the one-piece penguin under a transparent cover, and the one-piece penguin hidden under an opaque cover.

Based on looking time, Scott & Baillargeon conclude that infants expect the agent to search for the two-piece penguin where the one-piece penguin is, under the opaque cover, despite the fact that infants know the two-piece penguin to be under the transparent cover. The reason they give is that infants attribute to the agent the false belief that the penguin under the transparent cover is the one-piece penguin, since the agent did not see the two-piece penguin assembled to look like the one-piece penguin, as the infants did.

The problem with this interpretation is that, during familiarization trials, infants witness the agent, herself, assembling the two-piece penguin to look like the one-piece penguin. So they should attribute to the agent the belief that the two-piece penguin can be made to look like the one-piece penguin. But during the test trials infant interpreters completely fail to take this into account. This would be puzzling if they conceived of the agent as animated by an enduring mind, of which beliefs are potentially enduring states (Zawidzki 2011). Why would they ignore the extremely relevant fact that the agent's mind should be in the state of believing that the two-piece penguin can be assembled to look like the one-piece penguin?

If, however, these infants are employing a minimal theory of rational agency, in the sense described above, then this result is unsurprising. The infants do *not* think of their interpretive target as an agent animated by an enduring mind that enters different mental states of varying durations that are causally responsible for behavior. Rather, they think of their interpretive target as a bout of goal-directed behavior by a non-enduring agent, limited to the current spatiotemporal context, with certain relevant, transparent, and transient epistemic relations to present objects, like the two-piece penguin. Since this non-enduring agent fails to have direct line of sight on a relevant manipulation of the two-piece penguin, e.g., its assembly, and has a line of sight only on the assembled two-piece penguin, which looks one-piece, the infants conclude that the agent's goal-directed behavior will be misinformed.

In order to mount a general challenge to mentalist accounts of socio-cognitive capacities in infancy, we contrast our view with Butterfill and Apperly's (2013) account, which constitutes the most minimalist proposal still endorsing a mentalist interpretation of the data. On a first reading our proposal might seem very similar to theirs, which also explains infants' capacity to predict misinformed behavior in terms of their understanding of how situational invariants (e.g., the presence of barriers on line of sight between an agent and a possible target of action) may affect infants' habitual expectations about the future behavior of other agents. The difference between these two accounts, however, concerns their explanations of how infants form such expectations. Butterfill and Apperly are explicit that infant attributions of registrations count as attributions of mental states to agents, in virtue of the fact that registrations are intervening causal variables.[3] On our interpretation, in contrast, there is no attribution of an intervening causal variable to an agent. Rather, infants are attributing a relation between a bout of behavior and a (possibly non-actual) public situation—an attribution that can be updated when the infant notices specific situational invariants.

Moreover, we highlight a difference between the signature limits that Butterfill and Apperly identify for their minimal theory of mind, and the signature limits on our minimal theory of rational agency. The signature limits on a minimal theory of rational agency are supposed to distinguish *not* between the semantic properties of the states that infant and older interpreters attribute, as do Butterfill and Apperly's signature limits, but rather, between the respective social ontologies of infants and older interpreters. Adult human interpreters operating with a theory of mind conceive of the social world in terms of enduring agents animated by unobservable, enduring minds, states of which, like beliefs and preferences, are causally responsible for agent behaviors. On our view, instead, infant interpreters operating only with a minimal theory of rational agency conceive of the social world in terms of bouts of behavior with goals, performed by non-enduring agents that enter into an open-ended range of short-lived epistemic relations to items in environmental contexts the infants share with these agents. These short-lived epistemic relations determine whether or not the agents' behaviors are appropriately informed.

Given this way of distinguishing between infant and later interpretive competence, we should expect the following sorts of signature limits. (1) Infant interpreters should show no sensitivity to the effects on goal-directed behavior of *sufficiently dated* information, i.e., information that is not, in some sense, a component of the current interaction between the infant and the agent.[4] The reasoning here is that the only way to take such dated infor-

[3] "[T]heory of mind cognition begins when subjects ascribe *states* which function as *variables intervening* between environmental or behavioral inputs and behavioral outputs, and which play some roles characteristic of mental states (Whiten 1996; Penn and Povinelli 2007, 732). On this definition, *the endpoint in our construction (but no earlier point)* does count as theory of mind cognition because *registrations are intermediate variables and play a subset of the causal roles* characteristic of belief" (Butterfill and Apperly 2013, 621, emphasis added).

[4] We acknowledge that the notion of more or less dated information that may or may not be a component of an interaction is vague. However, this vagueness can be reduced by attending to the details of relevant experiments. For example, as we make clear in our dis-

mation into account is to conceive of the agent as animated by an enduring, unobservable mind, states of which causally explain the agent's sensitivity to spatiotemporally displaced situations. (2) Infant interpreters should be bad at binding goal-directedness and informedness to *particular* agents, since they do not conceive of these as products of causally potent states of enduring, unobservable minds within particular agents.

In fact, there is good evidence that infant interpretation is bound by these signature limits. Recent evidence from elicited-response, location change, "Sally-Anne" false belief tasks supports the first signature limit on infant sensitivity to relevant but *dated* information. Rubio-Fernández and Geurts (2013) analyzed the impact of two modifications to the traditional experimental setting on children's capacity to succeed in the task. First, they required children to *act out* the end of the story with the puppets that the experimenter already used to tell it, rather than asking them the final question about where Sally would look for the ball. Moreover, they also manipulated whether, at the point of the story when the main character leaves, the experimenter made her disappear completely by dropping the puppet under the table, thereby putting it out of the child's sight, or retained her on the scene with back turned. They found that even three-and-half-year-olds succeeded on the task when it was modified in *both* ways. In contrast, their success was below chance if either of the two modifications was not included.

It is known that the first modification, i.e., acting out the end of the story using the puppets rather than responding to a question, facilitates children's success in elicited response false belief tasks (Wellman et al. 2001), for at least two reasons. First, it increases children's engagement with the story, thereby enhancing their attention to the mental states of the characters involved—in particular, to the fact that Sally no longer knows where the ball is (Fenici submitted). Moreover, it removes a factor—i.e., mentioning the ball in the final question—that has been demonstrated to interfere with both children's (Rubio-Fernández and Geurts 2016) and even adults' (Rubio-Fernández and Geurts 2013) performance in the task.

The second modification is more crucial. If infants merely track bouts of behavior rather than the enduring mental states of other agents, they may fail to consider the persistence of the mental states of an agent who disappears from a scene, especially in social situations, like the experimental setting of the original elicited response false belief task, that pose significant linguistic demands. By eliminating the final verbal question, Rubio-Fernández and Geurts' first modification to the experimental paradigm reduces these

cussion of one such experiment below, infants seem to regard the complete disappearance of an agent from their sight as a disruption of the interaction, and hence fail to take into account to what situation the agent had epistemic access prior to its disappearance.

demands, thereby allowing children to rely on basic socio-cognitive capacities; presumably the same ones employed by younger infants in spontaneous false belief tasks. If these basic socio-cognitive capacities are limited to interpreting bouts of behavior by non-enduring agents, it is unsurprising that children continue to fail at the task when Sally disappears from the scene, yet succeed when she remains at the scene with back turned. If Sally disappears from the scene, to predict her behavior after her return, children would have to attribute mental states that endured while she was out of their sight, something that violates our first signature limit on the minimal theory of rational agency; so, on our view, we should expect children employing this competence to fail to properly complete Sally's behavior in this condition. In contrast, if Sally remains on the scene *without* an appropriate, observable epistemic relation to a relevant alteration, i.e., the object being hidden, children need only apply the minimal theory of rational agency to properly complete Sally's behavior; they need only conceive of Sally's behavior as *misinformed*, in the sense we specify above. In our view, this is why three-and-a-half year old children succeed *only with both* modifications of the original Sally-Anne task: subjects completing puppet behavior rather than responding to questions, and Sally remaining at the scene with back turned, rather than disappearing.

Evidence in favor of the second signature limit on the minimal theory of rational agency, i.e., inability to bind goals and information to particular agents, comes from an appropriate interpretation of another series of studies. It is known that infants do not interpret repeated reaching toward a location as intentional when no contrastive choice is available. For instance, infants expect an agent who repeatedly grasps or approaches an object to keep looking for that object only if, when the actor initially manifested her preference for the object, she also had the possibility to select a second object (Luo and Baillargeon 2005).

Exploiting this fact, Luo (2011; and, similarly, Luo and Baillargeon 2007, with slightly older infants) showed 10-month-olds an agent pushing an object beyond either a transparent or an opaque screen. A hand then appeared and removed it. This induced a false belief that the object was still there in the agent in the opaque condition, while in the transparent condition the agent could still see through the screen that the object was not there anymore. In both conditions, infants were then familiarized to see the agent reaching for another toy at a visible location. In the test trials, the position of the object was switched, and infants saw the agent reaching either for the old object at the new location or the new object at the old location.

The results showed that infants looked longer, indicating surprise, when they saw the agent reaching for the new object at the old location in the opaque, but not in the transparent, condition. This indicates that they had formed an expectation about the course of the observed action, and accordingly had interpreted the observed behavior as intentional, only when they considered that, from the perspective of the agent, there were two objects to choose between in the familiarization trials. According to Luo, it also suggests that "infants …considered the agent's informational states to decide whether or not to attribute a preference to her" (Luo 2011, 295).

Subsequent research however challenges this conclusion, and suggests instead the presence of a significant signature limit in infants' alleged capacity to attribute beliefs—a limit predicted by our present proposal. Kampis, Somogyi, Itakura, and Király (2013) ran the same experiment with a significant modification: in the test phase, they introduced a second agent who chose either consistently or inconsistently with the preference previously displayed by the first agent. Significantly, infants expected this second agent to choose as the first one, but only in the condition where this latter agent knew that two objects were present in familiarization.

This result confirms the empirical finding by Luo (2011) but also refines its significance. Indeed, if infants are simply tracking bouts of behavior, as we argued above, this result is exactly what we should expect: our second signature limit, i.e., infants using a minimal theory of rational action should be bad at binding goals and information to particular agents, implies that substituting one agent with another should not affect infant behavioral predictions. If, on the contrary, infants attribute enduring mental states to the first agent—as the advocates of mentalist interpretations of infant data argue—their action prediction capacities should be sensitive to *who* witnesses an event. Consequently, it remains a puzzle why they should attribute the very same mental state to a second agent who has never appeared before.

In light of these considerations, it is surprising that Kampis and colleagues maintain a mentalist interpretation of the data, and claim that these findings indicate that infants can attribute mental state contents (e.g., the belief content that there is such-and-such object behind the opaque screen), but do not bind them to agents. To whom are they attributing mental state contents?

In order to assess this mentalist interpretation, let us consider Kovács (2016), which defends a similar claim. Kovács reasons that, if infants attribute mental states such as beliefs, they should be able to individuate the beliefs they are attributing. However, "it seems that neither the agent nor the content alone could be sufficient to individuate a specific belief…. while …belief individuation or belief indexing should rely on a relation between the belief-holder and the belief-content" (Kovács 2016, 517). She thus introduces a theoretical construct, the "belief file", to "provide a representational structure with variables for (1) the agent, as the belief holder and for (2) the belief-content, in a way that each can be separately updated" (Kovács 2016, 515) but such that, together, they can be used to individuate a unique belief attributed to an agent.

Of course, Kovacs' interpretation is possibly true. We are not claiming that ours is the only possible interpretation. However, Kovacs' interpretation has the disadvantage of claiming that infants attribute mental states that have neither observable contents nor bearers. This is confusing because if the state has no bearer who might be misinformed about a situation, and it corresponds to no currently observable situation, what determines its content? And, furthermore, if it has no bearer, how can it be a *mental* content? Our alternative avoids these problems by denying that infants attribute mental contents. Rather, they attribute relations between bouts of behavior and information, based on observable, epistemic relations, which may involve agents other than the one currently being predicted. If infants are not attributing *mental* contents, then it is not strange to think that the epistemic relations of one temporally limited agent might affect the informedness of the behavior of a different temporally limited agent. It is true that this implies that infants attribute very complex, distributed, relational properties, e.g., a bout of behavior being misinformed by a non-actual situation in virtue of earlier, observable, epistemic relations to that situation when it was actual, by a different agent. But there is no reason to think infants aren't sensitive to such complex relational properties.

4. Discussion

While we think the hypothesis that infant interpreters employ a minimal theory of rational agency is promising, we are well aware that it is always possible to interpret the evidence discussed above in ways that salvage the hypothesis that infants attribute mental states. As they typically do in response to minimalist alternatives, defenders of this position can appeal to a strong competence/performance distinction. For example, infants might fail to take into account enduring mental states acquired in circumstances spatiotemporally displaced from the current context because of memory limitations. The idea is that they attribute enduring mental states to targets of interpretation, but cannot always remember all the mental states they have attributed. For instance, in Scott and Baillargeon's (2009) penguin experiment, during the familiarization trials, infants attribute to the agent the belief that the two-piece penguin can be assembled to look like the one-piece penguin, but they forget about this belief by the time they must interpret the agent during the test trial.

One problem with this idea is that infants appear to remember from familiarization that the agent desires to put the key in the two-piece penguin. So why do they remember one enduring mental state but not another? However, this enduring desire is a problem for our account as well: does it not show that infants *do* assume that some mental states endure?

From our perspective, the difference between the attribution of enduring goals and enduring belief states can be accommodated as follows. The infants do *not* think of the enduring goals as states of an unobservable, enduring mind that animates the agent; rather, they think of them in normative, agent-neutral terms: as kinds of behavior in which any agent ought to engage in the context of the penguin "game." On this view, rather than acquainting infant interpreters with the agent's enduring mental states, familiarization trials simply specify the rules of the game. This is consistent with Kampis et al.'s evidence that infants generally attribute goals in an agent neutral way.

Still, we acknowledge that the evidence seems interpretable either way. It is compatible with treating infants as scientific psychologists, with concepts of potentially enduring mental states that are causally responsible for behavior, yet with memory limitations. And it is also compatible with treating infants as detectors and parsers of spatiotemporally limited patterns of rational, goal-directed, (mis- or un-) informed actions, constrained by norms about what sorts of goals ought to be pursued in specific contexts.

Although we believe that future research will certainly provide additional data refining our understanding of both infants' socio-cognitive capacities and their limitations, we believe it is unlikely that the choice between these two alternatives will be settled by empirical research alone. These two views on social cognition in infancy derive from more general, alternative sets of interdependent assumptions belonging to the domain of "metatheories" rather than "theories" (Overton 2015). On the one hand, the mentalist interpretation of infants' social cognition privileges a view of mental state attribution capacities as extremely important in our social lives, and assumes that they are underpinned by dedicated neural processes (Saxe et al. 2004; Saxe and Wexler 2005), which develop under the pressure of cognitive maturation, and have been shaped through natural selection because of their survival value (Byrne and Whiten 1988; Humphrey 1976). In contrast, our proposal is grounded in the belief that mental state attribution capacities are not as central to our social lives as advocates of the mentalist interpretation assume. Far from being innately determined, they are progressively acquired by the child as she is introduced to the social practice of reporting others' mental states (Fenici forthcoming), and have evolved as responses to culturally evolved social practices, relative to which mental state talk was relevant to meeting social needs that have existed since the earliest human communities (Andrews 2012; Hutto 2008; Zawidzki 2013).

The contrast between these two general frameworks makes it unlikely that empirical evidence from a single experimental paradigm will decide the issue of which of them provides the more appropriate characterization of social cognition in infancy. It is more likely that the decision between such interpretations must be made on a broader, more theoretical basis. We believe that one important factor in deciding between them may come from assessing computational models implementing them (Pfeifer and Bongard 2006). Perhaps it is easier to implement, within a cognitive architecture, representations and predictions of bouts of goal-directed, (mis- or un-) informed behavior, by spatiotemporally limited agents than representations of states of enduring, unobservable minds. This might favor our interpretation. On the other hand, perhaps evidence from experiments involving adult social cognition will show that adults operate with a unitary socio-cognitive competence, rather than a hybrid, consisting of a minimal system conserved from infancy, and a later arriving sophisticated system. If this turns out to be the case, then it might support the mentalist interpretation of infant competence. If there is no trace of a distinct, minimal competence in adult social cognition, perhaps the more parsimonious hypothesis is that infants employ the same competence as adults, albeit in some attenuated form. Hence, if adult social cognition essentially involves concepts of mental states, then so does infant social cognition. In either case, the fact that resolving this dispute depends on such broad considerations suggests that the nature of infant social cognition is likely to be contested for a long time.

5. Conclusions

Do infant socio-cognitive abilities manifested in spontaneous response false belief tasks indicate a capacity to attribute unobservable, enduring mental states? Challenging a widely shared assumption among cognitive scientists and developmental psychologists, we have proposed that, rather than attributing mental states to agents they interpret, infants merely track goals of observed bouts of behavior, as well as the situational invariants affecting whether or not such behaviors are appropriately informed. We believe that we have provided good reasons to explore alternative ways of interpreting the experimental data coming from studies of infant social cognition, though we have not provided a decisive argument against prevailing mentalist interpretations. After all, the debate between mentalist and non-mentalist theories of social cognition is likely driven by highly theoretical, background assumptions; so, it is unlikely that results from a relatively narrow experimental paradigm will conclusively favor either interpretation of infant socio-cognitive competence.

Bibliography

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

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

Baillargeon, R., Scott, R. M. and He, Z. (2010). False-belief understanding in infants, *Trends in Cognitive Sciences* **14**: 110–118.

Baron-Cohen, S., Leslie, A. M. and Frith, U. (1985). Does the autistic child have a "Theory of Mind"?, *Cognition* **21**: 37–46.

Bloom, P. and German, T. P. (2000). Two reasons to abandon the false belief task as a test of Theory of Mind, *Cognition* **77**: 25–31.

Buckner, C. (2014). The semantic problem(s) with research on animal mind-reading, *Mind & Language* **29**: 566–589.

Butterfill, S. A. and Apperly, I. A. (2013). How to construct a minimal theory of mind, *Mind & Language* **28**: 606–637.

Byrne, R. W. and Whiten, A. (eds) (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*, Oxford University Press, Oxford.

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

Csibra, G., Bíró, S., Koós, O. and Gergely, G. (2003). One-year-old infants use teleological representations of actions productively, *Cognitive Science* **27**: 111–133.

Csibra, G., Gergely, G., Bíró, S., Koós, O. and Brockbank, M. (1999). Goal attribution without agency cues: The perception of "pure reason" in infancy, *Cognition* **72**: 237–267.

Fenici, M. (2014). A simple explanation of apparent early mindreading: Infants' sensitivity to goals and gaze direction, *Phenomenology and the Cognitive Sciences* **14**: 1–19.

Fenici, M. (2015). Social cognitive abilities in infancy: Is mindreading the best explanation?, *Philosophical Psychology* **28**: 387–411.

Fenici, M. (forthcoming). What is the role of experience in children's success in the false belief test: Maturation, facilitation, attunement, or induction?, *Mind & Language*.

Fenici, M. (submitted). How children approach the false belief test: Social development, pragmatics, and the assembly of theory of mind.

Fenici, M. and Carpendale, J. I. M. (submitted). Overcoming the false belief test puzzle: A constructivist approach to the development of social understanding.

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

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

Helming, K. A., Strickland, B. and Jacob, P. (2016). Solving the puzzle about early belief-ascription, *Mind & Language* **31**: 438–469.

Heyes, C. M. (2014). False belief in infancy: A fresh look, *Developmental Science* **17**: 647–659.

Humphrey, N. K. (1976). The social function of intellect, *in* P. P. G. Bateson and J. R. Hinde (eds), *Growing Points in Ethology*, Cambridge University Press, Cambridge, pp. 303–317.

Hutto, D. D. (2008). *Folk Psychological Narratives*, The MIT Press, Cambridge, MA.

Hutto, D. D. (2015). Basic social cognition without mindreading: Minding minds without attributing contents, *Synthese* **Online first**: 1–20.

Kampis, D., Somogyi, E., Itakura, S. and Király, I. (2013). Do infants bind mental states to agents?, *Cognition* **129**: 232–240.

Kovács, Á. M. (2016). Belief files in theory of mind reasoning, *Review of Philosophy and Psychology* **7**: 509–527.

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

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

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

Luo, Y. and Baillargeon, R. (2010). Toward a mentalistic account of early psychological reasoning, *Current Directions in Psychological Science* **19**: 301–307.

Nichols, S. and Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford University Press, New York.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?, *Science* **308**: 255–258.

Overton, W. F. (2015). Processes, relations, and relational-developmental-systems, *in* W. F. Overton and P. C. M. Molenaar (eds), *Handbook of Child Psychology and Developmental Science*, pp. 9–62.

Penn, D. C. and Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **362**: 731–744.

Pfeifer, R. and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*, The MIT Press, Cambridge, MA.

Povinelli, D. J. and Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind, *Mind & Language* **19**: 1–28.

Rubio-Fernández, P. and Geurts, B. (2013). How to pass the false-belief task before your fourth birthday, *Psychological Science* **24**: 27–33.

Rubio-Fernández, P. and Geurts, B. (2016). Don't mention the marble! The role of attentional processes in false-belief tasks, *Review of Philosophy and Psychology* **7**: 835–850.

Ruffman, T. (2014). To belief or not belief: Children's theory of mind, *Developmental Review* **34**: 265–293.

Ruffman, T. and Perner, J. (2005). Do infants really understand false belief? Response to Leslie, *Trends in Cognitive Sciences* **9**: 462–463.

Saxe, R., Carey, S. and Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging, *Annual Review of Psychology* **55**: 87–124.

Saxe, R. and Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction, *Neuropsychologia* **43**: 1391–1399.

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

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

Southgate, V., Senju, A. and Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds, *Psychological Science* **18**: 587–592.

Träuble, B., Marinović, V. and Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs?, *Infancy* **15**: 434–444.

Wellman, H. M., Cross, D. and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief, *Child Development* **72**: 655–684.

Whiten, A. (1996). When does smart behaviour-reading become mind-reading?, *in* P. Carruthers and P. K. Smith (eds), *Theories of Theories of Mind*, Cambridge University Press, Cambridge, pp. 277–292.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception, *Cognition* **13**: 103–128.

Zawidzki, T. W. (2011). How to interpret infant socio-cognitive competence, *Review of Philosophy and Psychology* **2**: 483–497.

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*, The MIT Press, Cambridge, MA.