Synthese (2017) 194:673–688
DOI 10.1007/s11229-015-0792-3

S.I.: FUTURE OF SOCIAL COGNITION

Mindreading in adults: evaluating two-systems views

Peter Carruthers[1]

Received: 13 December 2014 / Accepted: 2 June 2015 / Published online: 23 June 2015

Abstract A number of convergent recent findings with adults have been interpreted as evidence of the existence of two distinct systems for mindreading that draw on separate conceptual resources: one that is fast, automatic, and inflexible; and one that is slower, controlled, and flexible. The present article argues that these findings admit of a more parsimonious explanation. This is that there is a single set of concepts made available by a mindreading system that operates automatically where it can, but which frequently needs to function together with domain-specific executive procedures (such as visually rotating an image to figure out what someone else can see) as well as domain-general resources (including both long-term and working memory). This view, too, can be described as a two-systems account. But in this case one of the systems encompasses the other, and the conceptual resources available to each are the same.

Keywords Automatic · Executive function · False belief · Mindreading · Perspective taking · Two systems

1 Introduction

Throughout most of the first twenty-five years of research on human mindreading capacities (or "theory of mind," as it is often called), the focus was on questions of development. It was generally assumed that adults possess a single mindreading system. The question was how that adult competence is acquired, and via what intermediate stages. A widespread consensus emerged that adult-like competence (specifically an understanding of belief, false belief, and misleading appearances) only appears in

✉ Peter Carruthers
pcarruth@umd.edu

[1] Department of Philosophy, University of Maryland, College Park, MD 20742, USA

children around the age of four (Wellman et al. 2001), although everyone allowed that learning continues to take place thereafter. Most of the tasks employed in these studies used verbally presented materials and/or verbal responses to questions, however. And there were always some who insisted that these verbal tasks might mask an earlier underlying mindreading competence (Leslie et al. 2004). More recently there have been a flurry of findings using implicit (non-verbal) tasks suggesting that infants between the ages of 6 and 18 months can reason appropriately about the false beliefs and subjective experiences of others (Onishi and Baillargeon 2005; Southgate et al. 2007; Surian et al. 2007; Song and Baillargeon 2008; Song et al. 2008; Buttelmann et al. 2009; Poulin-Dubois and Chow 2009; Scott and Baillargeon 2009; Kovács et al. 2010; Scott et al. 2010; Southgate et al. 2010; Träuble et al. 2010; Luo 2011; Senju et al. 2011; Knudsen and Liszkowski 2012; Yott and Poulin-Dubois 2012; Baillargeon et al. 2013; Barrett et al. 2013; Buttelmann et al. 2014; Southgate and Vernetti 2014; Buttelmann et al. 2015).

Interpretations of these recent infancy findings have occupied all points along the spectrum. Some have denied that they reveal mindreading competence of any kind. Rather, it is said that the data can be explained by appeal to low-level stimulus features and associations (Perner and Ruffman 2005; Heyes 2014) or mere behavior-rules of some sort (Perner 2010; Gallagher and Povinelli 2012). Others have insisted that the data demonstrate the existence in infants of sophisticated mindreading competence of an adult-like kind, and have appealed to performance factors to explain infants' failures in verbally mediated tasks (Baillargeon et al. 2010; Carruthers 2013). Yet others, however, have adopted an intermediate position. While allowing that infants are capable of mentalizing of a sort, they have denied that infants represent beliefs and desires *as such*. Rather, infants employ a simpler set of concepts housed in a fast and efficient mental-state-tracking system. This system is encapsulated from verbal report (hence toddlers' failures in verbal tasks), and continues to operate automatically in older children and adults (Apperly and Butterfill 2009; Apperly 2011; De Bruin and Newen 2012; Butterfill and Apperly 2013).

Relatedly, and in part inspired by these recent infancy results, researchers have also turned to investigate mindreading capacities in adults. Many of these latter findings have been interpreted as supporting the existence of two distinct (or largely distinct) mindreading systems: one that is fast, automatic, but inflexible, and the other of which is slower, controlled, and yet flexible (Keysar et al. 2003; Apperly et al. 2006; Back and Apperly 2010; Kovács et al. 2010, 2014; Lin et al. 2010; Qureshi et al. 2010; Samson et al. 2010; Surtees and Apperly 2012; Schneider et al. 2012a, b, 2014a; Surtees et al. 2012; Low and Watts 2013; van der Wel et al. 2014). When taken together with the recent infancy data showing mindreading competence in infants using implicit tasks, these findings have been interpreted not just as evidence of the existence of two sets of mindreading *procedures*, but as providing support for systems that differ from one another in their conceptual resources (Apperly and Butterfill 2009; Apperly 2011; Butterfill and Apperly 2013; Low et al. 2014). The idea is that there is one system that is available in infancy and that continues largely unchanged into adulthood. This system employs a set of concepts that enables it to track mental states in certain circumstances without being able to represent them as such, and while being encapsulated from verbal report. But there is, in addition, another system that is constructed slowly over time, relying partly on linguistic input, and in which our familiar adult concepts of belief, desire, and subjective experience find their home.

I have discussed the infancy data at length elsewhere (Carruthers 2015), arguing that they provide little support for a two-systems view. Rather, the evidence can be better explained by postulating a single system that represents mental states as such from early in infancy, and that is gradually elaborated through learning and slowly integrated with the child's executive systems and linguistic abilities. This early system *becomes* the adult system over time. On this view, the reason why toddlers fail explicit mindreading tasks is not for lack of conceptual competence. Rather it is because such tasks are, in effect, triple-mindreading ones. The child must devote mindreading resources to interpreting the speech of the experimenter and then formulating a communicative response while at the same time representing and drawing inferences from the mental states of the target agent. It makes sense that this might overwhelm young children's abilities. For their mindreading, executive, and language systems, as well as the connections between them, are not yet fully mature.[1] Consistent with this interpretation, when 2.5-year-olds are tested for false-belief understanding using verbally-presented stories combined with an implicit outcome measure (thus removing the need for the mindreading system to participate in formulating a communicative response), they pass (Scott et al. 2012).

If the infancy data fail to provide support for a two-systems account (as I propose to take for granted hereafter), then the entire weight of the argument for such a view must fall on the recent findings with adults. The present discussion will confront the adult data directly, arguing that they, too, fail to support a two-systems account of the sort sketched above. One set of data concerns people's inferences about the visual access and experience of others; another set concerns people's inferences about belief and false belief. We will consider these in Sects. 2 and 3 respectively, before elaborating the proposed one-system explanation in Sect. 4. It should be stressed, however, that the present paper is not intended to convince the reader of the correctness of the latter. To do that, one would also need a full examination of the infancy and child data. Rather, the goal of the present paper is to convince the reader that the adult data provide no independent *support* for a two-systems view, and that those data (considered in isolation) are more parsimoniously explained by a one-system account.

One final set of preliminary comments before we begin our main discussion concerns the meaning of "automatic" and related terms. I shall understand an automatic process to be one that takes place independently of the agent's goals (whether implicit or explicit). Note that this need not imply that the process is a *mandatory* one, however, meaning that it cannot be *inhibited* by the subject (although many automatic processes are, no doubt, also mandatory). Automatic processes should also be contrasted with *spontaneous* ones. These take place independently of external prompting and explicit (conscious) goals, but they nevertheless depend on implicit goals and hence require executive resources. As we will see, some of the mindreading processes investigated in adults seem to be genuinely automatic whereas others are merely spontaneous.

[1] The explanation sketched here is similar to that provided by Baillargeon et al. (2010), except that it emphasizes the three-way connection between language, executive function, and mindreading, and not just among the latter two.

2 Seeing what others see

There is evidence of a system that automatically and inflexibly computes the visual access of others. For people will encode what another agent can see even in circumstances where this is irrelevant to the person's own task, and indeed even if it *impedes* the person in their own task (Qureshi et al. 2010; Samson et al. 2010; Surtees and Apperly 2012). In these experiments people have to respond, as quickly and accurately as possible, to say whether they themselves can see one dot or two in a visual display, or whether an avatar within the display can see one dot or two. On each trial they are told whether they are to judge what they can see, or what the avatar can see. Naturally, and as one might expect, people are somewhat slower and make significantly more errors when judging what the avatar can see in circumstances where this conflicts with what they themselves can see. For example, they might mistakenly respond "two" in circumstances where the avatar can only see one dot, but where they themselves can see two. The interesting finding, however, is that there is just as much interference in the reverse direction. Subjects who merely have to report what they themselves can see are nevertheless slower and make more errors when the avatar can see something different. It seems that people cannot help but encode what the avatar sees, and this representation then competes for control of their motor response, either slowing down a correct response or substituting an incorrect one.

There are other circumstances in which people do *not* display interference from what another person can see, however, suggesting that the visual states of another agent have *not* been automatically represented (Surtees et al. 2012). These experiments employ a similar avatar-involving situation to that described above, but where computation of what the other person can see requires "level-2 perspective taking" (Flavell 1978). As before, people are cued on each trial to say either what they or the avatar can see. But now, instead of making judgments of one dot versus two, they are judging whether they, or the avatar, can see a specific Arabic numeral which is situated on the desk or the wall in between the subject and the avatar (who sits facing the subject). Some of these numerals (like the numeral "8") are symmetrical around the horizontal axis, and so would appear the same to both the subject and the avatar. But others ("6" and "9" in particular) will appear as one numeral to one person ("6", say) and as another to the other ("9"). In these circumstances people are *not* slower or less accurate in judging that they themselves can see a "6", suggesting that they have not automatically encoded that the avatar can see a "9". It has been claimed, as a result, that a signature limit of the automatic system is that it represents what *things* another agent sees without encoding the *aspect under which* the agent sees them (Butterfill and Apperly 2013). The latter is said to require the resources of a distinct mindreading system that is non-automatic, effortful, and executively controlled.

These finding are more parsimoniously explained, however, by supposing that there is a single mindreading system that computes what other agents can see, which employs the same concept of *seeing* across conditions, and which makes these attributions automatically where it can do so (where it does not need to draw on additional executive resources). In circumstances that require such resources, however, attributions are task-dependent, and are only computed if required. On this account, the difference between the two sets of experiments reviewed above is that in the first set the mindreading system only needs to compute line-of-sight to represent what the avatar can see. In effect, what the avatar can see can be taken in at a glance (given the salience of the dots on the wall to the ongoing task as a whole), and is automatically represented. In order to determine whether the avatar sees “6” or “9”, in contrast, the participant must start from an image of what they themselves see (“6”, say) and rotate it through 180 degrees to generate an image of what the avatar can see, which can then be recognized as a “9”. Since this is an executively demanding task, it is only performed when it is relevant to the participant’s primary goal (that is to say, when the participant is cued to say what the avatar can see). Note that there is nothing in this account to require that the concept *see* in “The avatar can see two dots” is any different from the concept employed in “The avatar sees ‘nine’.”

Ironically, data from the same lab directly supports this simpler interpretation (Surtees et al. 2013). In these experiments participants make judgments of both visual *and* spatial perspectives. All of the stimuli contain an avatar sitting in a room with a large numeral on the floor nearby. In some trials participants have to judge whether the numeral is in front of or behind the avatar, or whether it is to his left or right. In other trials participants have to judge whether the avatar can see a specific numeral. Some of these, like “8”, can be identified independently of perspective (provided it is in the avatar’s line of sight). Others are ambiguous, and would differ depending on perspective, like “6” and “9”. Across all trials what varies is the avatar’s perspective with respect to the observer: in some he sits facing the observer and in some he sits with his back to the observer, as well as in a variety of other orientations.

The main finding is that the avatar’s orientation has no effect on either speed or accuracy when participants make judgments of front / behind, nor when they judge unambiguous instances of see / not-see. In contrast, *both* in the case where participants judge left / right *and* where they judge “6” / “9”, response times and error rates increase as a function of the avatar’s orientation. (Judgments are easier when the participant and avatar face in the same direction and most difficult when the avatar sits facing the participant.) The best explanation, then, is that in these versions of the tasks participants are visually rotating their own image to match the perspective of the avatar, with larger angular rotations taking longer, as we know from image-rotation experiments generally (Kosslyn 1994). Moreover, there is no reason here to think that distinct concepts of *see* are employed when judging “Sees ‘eight”’ and when judging “Sees ‘six”’.

A similar resource-based explanation can be given of the other main body of findings suggesting that people do not automatically encode the visual perspective of another agent (Keysar et al. 2000). In these experiments an addressee and a communicator sit on opposite sides of an array of shelves containing a variety of objects. Some of the objects are mutually visible, but some are not, because the side of the shelf facing the communicator is blocked off. (Participants are thoroughly familiarized with the properties of the shelves, sometimes by playing the role of the communicator in an initial phase of the experiment). Addressees might hear an instruction like, “Move the small candle up one shelf” in circumstances where there are a number of candles visible, the smallest of which only they can see. In such cases they nevertheless show more fixations on that candle initially, and often initiate a movement in the direction of that candle (and sometimes actually move it). It seems that participants are not automatically computing what the communicator can or cannot see, but are only figuring this out after the instruction is given, correcting their own initial egocentric response accordingly. Later experiments have demonstrated that successful interpretation of the communicator's instruction in these conditions depends on the resources of working memory, since people with low working memory take longer to reach for the correct object and make more errors, but only in cases of perspective-conflict. Moreover, people placed under working-memory load perform likewise (Lin et al. 2010).

These experiments, too, do nothing to support the existence of two distinct systems for mindreading. For no one should think that an automatically-operating mindreading system would attempt to encode *everything* that an agent can see (and hear, and feel, and smell, etc.). Rather, the most plausible hypothesis is that it would only encode events as perceived by an agent that are perceptually accessible to the target agent *and* are salient to the encoder in some way. So one would not predict that an automatically-operating mindreading system would attempt to build a detailed model of everything that the communicator can and cannot see in the experimental setup. Instead, attributions would be made on the fly, cued by direction of gaze to salient objects (if eye-direction is visible), or by speech that uniquely picks out a given object. A model of the agent's awareness would thus be built up incrementally by actions that demonstrate awareness of specific objects. Consistent with these suggestions, Hanna et al. (2003) find that when communicators have previously referred to a specific mutually-visible object, participants no longer look first toward the competitor object that only they themselves can see (as they do when the target object has not previously been referred to).

I conclude that there is nothing on the data concerning adults' attributions of *seeing* to other agents to suggest the existence of two distinct mindreading systems that contain different concepts of *see*. Rather, the data are better explained in terms of a single mindreading system that automatically attributes mental states to others when the computations involved are straightforward, but which needs to co-opt the resources of executively-controlled working-memory procedures in a task-dependent manner in cases where the computations required are more demanding.

3 Seeing what others believe

There are now a number of bodies of evidence suggesting that adults will automatically track the beliefs and false beliefs of another agent in some circumstances. Kovács et al. (2010) required participants to watch a simple animated display in which a ball moved behind a screen, sometimes remaining there, sometimes re-emerging only to go behind the screen again, sometimes emerging to leave the scene altogether, although sometimes re-appearing from off-stage to end up behind the screen. Throughout some or all of the sequence an avatar watched from the side of the stage, playing no part in the proceedings, and being irrelevant to the participants' task, which was to press a button as swiftly as possible if a ball was present behind the screen when it dropped at the end of the sequence. Note that the putative beliefs of the avatar about the location of the ball would sometimes differ from those of the participant. On some trials a ball was unexpectedly present when the screen dropped, although the participant had seen it leave the stage. Participants were slower to respond when both they and the avatar expected that there would not be a ball behind the screen (although there was). And naturally, participants were significantly faster when the presence of the ball was expected. But their responses were speeded just as much on trials where they themselves did not expect the ball to be there but the avatar did. It seems that participants were automatically encoding what the avatar thought about the location of the ball, and that this was priming their own responses.

These data have been criticized by Phillips et al. (2015), who argue that they are an unintended artifact of the timing of the periodic "attention checks" that were administered throughout the procedure. (Participants were required to press a button to record that they had noted the presence of the avatar.) Phillips and colleagues argue that when the timing of the attention checks vary, but beliefs do not, the effect is present; whereas when beliefs vary while the attention checks are held constant, the effect is absent.[2] On the other hand, van der Wel et al. (2014) provide a conceptual replication of the original findings using a continuous measure. Their stimuli involved a ball and two screens, rather than one, and the task was to *reach* as swiftly as possible to the location of the ball when both screens dropped. In some conditions participants believed that the ball would be behind one screen whereas the avatar believed that it was behind the other. In these circumstances participants' reaching motions showed an influence of the avatar's belief, following a trajectory that deviated in the direction of the location expected by the avatar.[3] Moreover, Kovács et al. (2014) provide a brain-imaging replication of the original finding, showing that a crucial component of the mindreading network (the right temporo-parietal junction) is active in cases where the incidental avatar holds a differing belief about the location of the object.

Other experiments have used eye-tracking methods to demonstrate the presence of false-belief representations that are entirely peripheral to the participants' task (Schneider et al. 2012a). These tests adapted the procedure previously used with infants and adults by Southgate et al. (2007) and Senju et al. (2010).[4] Participants watch videos in which a puppet transfers a ball from one box to another while an agent watches or (in some conditions) has left the room. Their task is to press the spacebar every time the agent waves at the camera, so they have no motivation to track the agent's mental states. But they have learned that when a bell sounds the agent will reach for one or other of the boxes to retrieve the ball. The participants' anticipatory eye movements demonstrate that they track the beliefs of the agent and form expectations accordingly. Yet in-depth follow-up interviews show that few participants have any conscious awareness of doing so. Moreover, Schneider et al. (2014a) not only replicate

[2] I gather that a response to these criticisms is forthcoming from Kovács and colleagues. They point out that the printed feedback at the bottom of the screen that accompanied the attention-checks used by Phillips et al. (2015) proved highly distracting, to the point where the resulting data failed even to show any effect of the participants *own* beliefs. Small wonder, then, that there should have been no effect of the avatar's beliefs either, in these circumstances.

[3] In contrast with the Kovács et al. (2010) findings, the effect of the avatar's belief is much smaller than the agent's own when analyzed using this continuous measure, which could also explain why Phillips et al. (2015) were unable to find any evidence of it using a discrete measure.

[4] While these earlier methods likewise demonstrated task-independent belief-based action prediction, there was no accompanying primary task and no measure of subjects' awareness. Hence although belief-tracking was task-independent, it is possible that it was nevertheless deliberate and hence neither automatic nor spontaneous.

the effect, but show that these anticipatory eye movements occur independently of task instructions, occurring in the same manner whether people are instructed to predict what the agent will do, or are instructed to keep track of the location of the ball, or are given no instructions.

In contrast with this evidence of automatic false-belief attribution, other studies suggest that such reasoning is *not* automatic. One of these can be dealt with in the same manner used to critique the visual-rotation-involving perception tasks in Sect. 2. This is Low and Watts (2013). This study, too, is modeled on the procedure followed by Southgate et al. (2007), and uses eye-tracking to measure anticipatory looking. Children and adults are first familiarized with the fact that the target agent likes blue things rather than red things, and that she will reach through the appropriate door to retrieve the desired blue object shortly after the doors light up and flash at the end of the trial. Then they watch as the agent observes what appears to the participants to be a red object move from one box to the other. Then out of sight of the agent the object rotates back and forth to reveal that while it is red on one side it is blue on the other; hence the *agent* would have seen a *blue* (and hence desirable) object enter the box. The object then leaves that box and returns to the previous one, this time with its blue side facing the participant. Thus the target agent, seeing what she would take to be a red object leave the box, should think that the blue object is still there. Belief-reasoning should therefore lead participants to expect that the agent will reach for the now-empty box when the doors flash. However, no group of participants show anticipatory looking suggestive of belief reasoning. Yet both 4-year-olds and adults give correct answers to explicit questions when asked to predict the agent's action. This leads the experimenters to propose that an inability to represent the *aspectual* nature of belief is a signature limit of the automatic mindreading system.

It is surely plain, however, that this task is not one that could be executed automatically, without making motivated use of executive function and working memory. Indeed, if we do suppose that belief attribution is automatic, the belief that would automatically have been encoded for the agent initially is that she thinks a *red* object has entered the box. In order to update this attribution when participants discover the double-sided nature of the object, they would then need to access a memory of the initial event and visually rotate the moving red object to figure out that the agent would have seen a *blue* object move across the stage. Then once again when the object leaves the box, visual rotation would be required to infer that the agent would have seen a red object leaving. What these data really demonstrate, therefore, is just that in some circumstances mindreading resources need to work in conjunction with executively controlled uses of long-term and working memory. That this sort of executive control does not happen automatically should be no surprise to anyone. Indeed, non-automaticity is often *defined* in terms of executive control.

Another body of data suggesting that belief-reasoning is not automatic is provided by Apperly et al. (2006) and Back and Apperly (2010). In these experiments participants watch while a male agent plays a kind of "shell game" with a ball, placing it first under one cup and then under another. The participants' task is to keep track of the location of the ball, and to indicate its current location as quickly as possible when probed to do so. Meanwhile a female agent watches some or all of the proceedings, leaving the room at some points and returning at others. On some occasions participants are unexpectedly probed about the female agent's beliefs. Reaction times are slower for these reports than for reports of the object's location, both for true and for false beliefs. Yet when participants are instructed to keep track of the agent's belief this differential disappears. (Indeed, in some conditions belief-reports are faster.) This suggests that belief-reporting is not intrinsically harder than location-reporting. In interpretation of their data, the experimenters suggest that participants do not attribute beliefs to agents automatically, but only compute them in response to task demands.

Notice, however, that there is generally a significant lapse of time between the point at which the female agent would have formed her most recent belief about the object's location (before she left the room) and the later belief probe. So the experiment is not really about *encoding* belief but *recalling* it. Supposing that the agent's beliefs were automatically tracked and encoded as the scenario unfolds, there are just two ways in which the relevant information could be made available to answer the unexpected belief-probe. Either that information would need to be maintained actively in working memory, or it would need to be retrieved from long-term memory in response to the probe. We know that it isn't held in working memory, because of the extra response time. And why would it be? For participants think their task is to track the object's location, not the female agent's beliefs. Hence the information about the agent's beliefs (supposing that it had been automatically encoded) would need to be retrieved from memory when participants are unexpectedly probed. Naturally it takes longer to access long-term memories than it does to respond on the basis of information already held in working memory (in the way that information about the ball's actual location surely is).

The difference in response times found by Apperly and colleagues can therefore be explained while supposing that people *do* automatically encode the beliefs of others (where these are salient and people can do so easily). Consistently with this account, Cohen and German (2009) use a version of the same procedure and show that when the belief-probes occur much closer in time to the events that signal the female agent's belief, responses are faster than responses about reality and just as fast as responses to belief-probes when subjects are instructed to track beliefs. Note that in these circumstances we can assume that a representation of the agent's belief would still be readily available, and hence would not need to be searched for in long-term memory.

I conclude that there is nothing in the data considered here to challenge the idea that people automatically encode the beliefs of salient agents about salient events. Apparent non-automaticity may derive either from the costs of drawing inferences about already-encoded beliefs (in some situations) or from the executive demands of retrieving previously-encoded beliefs from long-term memory (in other situations). So there is nothing here to support the existence of two distinct systems for belief-attribution that contain differing concepts of belief. These claims would be strengthened, however, if we could at least sketch a general model of how an automatically-functioning mindreading system operates and interacts with executive systems, drawing on long-term memories and the resources of working memory when needed. That will be the task of the next section.

4 Automatic and not

What single-system model would best explain the pattern of results discussed above? We can suppose that there is a single mindreading system containing concepts of *belief*, *desire*, and *see*. This system is engaged whenever an agent is identified as such, and it automatically tracks what the agent can see, using cues such as eye-direction and the saliency of objects and events in the agent's line of sight. These attributions, in turn, automatically give rise to attributions of the corresponding beliefs. (That is, the system transitions automatically from, *The agent sees the ball going into the box*, to, *The agent thinks the ball is in the box*. This will become an attribution of a belief that is false if the ball is moved again while the agent is absent, since this representation will not then be updated).[5] At the same time the system automatically tracks what the agent wants, where this can be inferred easily from patterns of goal-directed movement or other cues (facial expressions, verbal statements, and so on). In effect, we can suggest that the mindreading system automatically builds a partial model of the mental states of any agent that it encounters, provided that little or no executive control is required beyond the allocation of attention to relevant aspects of the stimuli. We can suppose that the various components of this mental-state model reverberate actively for a short while after they have been created, before subsiding into long-term memory.[6]

But what of behavior prediction? What of the use of mental-state information to predict what an agent will do? We can suppose that this, too, is fully automatic where the actions in question can be predicted from goal-states and perception-states that have just been attributed to the agent, and whose representations will therefore still be active. We can suppose, indeed, that these predictions show up as so-called "mirror neuron" activity in premotor cortex. This is an interpretation of mirror-neuron activity that has been gaining increasing ground among theorists recently. The suggestion is that the main role of mirror neurons is predictive, driven by goals attributed to the agent on the basis of features of the context together with initial behavior (Csibra 2007; Jacob 2008).

[5] Note that on this account no special representational resources are required to represent a false belief. Truth and falsity can remain implicit in the procedures for updating, or for not updating, an agent's belief when circumstances change while the agent is either present or absent. And it may be that infants, too, represent true and false beliefs without predicating truth and falsity as such. (The latter concepts can at some point be introduced through a pair of straightforward definitions: S has a true belief that P = P & S believes that P; S has a false belief that P = not-P & S believes that P. No radical conceptual change—in the sense of Carey (2009)—is therefore required.) This might lead some people to deny that infants are representing beliefs *as such*. It might be claimed that only when children acquire concepts of *truth* and *falsity* and come to understand explicitly that beliefs can be false do they really represent beliefs as such. But as Carruthers (2015) argues at length, this sort of maneuver is merely definitional, and does nothing to support a two-systems account. For it can be the very same representations that are present in infancy that are gradually elaborated and enriched as the child learns more and more about the mind (and acquires explicit concepts of truth and falsity). And then in adulthood, too, people may tap into more or less of this rich body of knowledge depending on task demands. It can still be the same belief-representation that figures both in implicit tasks of the sort discussed in Sect. 3 and also in Sherlock–Holmes-type cases where one reflects consciously about someone's false beliefs while drawing on a much richer body of background knowledge.

[6] Whether they get *consolidated* in long-term memory or slowly fade away—hence belonging to what is sometimes called "long-term working memory" (Ericsson and Kintsch 1995)—is another matter, of course, depending on levels of emotional engagement and other factors.

On this account, people automatically anticipate what other agents will do next in light of what they take to be their current goals.

In many of the cases of putatively-automatic behavior prediction discussed earlier, however, the goals and beliefs necessary to predict behavior will have been computed some time previously. There is good reason to think that they are not actively maintained in working memory in the interim, since the contents of working memory are generally agreed to be conscious, and yet Schneider et al. (2012a, 2014a) could find no evidence of conscious awareness of mentalizing in their experiments. We can therefore conclude that the relevant components of the model will need to be retrieved from long-term memory. This is a function generally attributed to the executive. But we can suppose that people have a standing or habitually-formed goal of anticipating the behavior of other agents, especially where they know from previous experience that an action of some sort is about to be undertaken. (Recall that in Schneider et al.'s experiments the doors flash to signal that the agent is about to reach through one of them to obtain the target object.) The result is that long-term memories of the target agent's mental states are searched for and activated, and then used to issue a prediction. (All this happens outside of awareness, of course).

The upshot of this account is that unconscious mindreading is not always strictly automatic; not, at any rate, if "automatic" implies "goal-independent". It can nevertheless still be spontaneous, in the sense of being independent of any overt task or consciously-accessible goal. We know that mindreading can take place in the absence of any externally-induced goals, since people's anticipatory looking is the same whether they are asked to track the target's mental states, or the location of the ball, or are given no instructions (Schneider et al. 2014a). And we know that it can take place unconsciously, as we have just noted. But these facts are consistent with the standing-goal account sketched above, since there are good reasons to think that goals can be active and do their work outside of people's awareness (Dijksterhuis and Aarts 2010; Marien et al. 2012; Huang and Bargh 2014). Moreover, the standing-goal account can explain why apparently-automatic mindreading should nevertheless collapse under executive load (Schneider et al. 2012b). For we can suppose that the behavior-predicting goal is down-regulated in importance or suppressed altogether when people simultaneously undertake another, executively demanding, task.

The same account can also explain *failures* of automatic mindreading, of course. If what another agent sees has not been computed in advance (because it isn't salient until a verbal request makes it so, as in the experiments of Keysar et al. 2000), then that information will need to be computed after the fact, leading to hesitation, delay, and sometimes outright error. Likewise if computing what another agent sees requires executively-controlled rotation of a visual image in working memory (as in the experiments of Surtees et al. 2012), then this is unlikely to be done without specific motivation to do so (such as compliance with an experimental request, as happens in the *other-sees-"9"*-condition). And for the same reason, one will not automatically compute what an agent believes when it would require working-memory-involving inferences to do so (as in the experiments of Low and Watts 2013). Moreover, even if another agent's beliefs *have* been automatically computed, the resulting representations will need to be accessed from long-term memory after more than a few seconds have elapsed (as in the experiments of Apperly et al. 2006; Back and Apperly 2010), leading to apparent failures of automaticity.[7]

On the account outlined here there is just a single mindreading system housing a single set of mental-state concepts, attribution procedures, and prediction strategies. This system can operate fully automatically in simple cases, or it can operate spontaneously (but still unconsciously) in others when combined with a standing goal of predicting behavior (needed to initiate searches of long-term memory). But it can also operate together with controlled uses of working memory in tasks that require visual rotation. Moreover, we can assume that it is this same system whose resources contribute to fully explicit ("Sherlock Holmes style") reasoning about the mental states of others, which might utilize sequences of inner speech and other forms of working-memory activity as one slowly and reflectively tries to figure out what someone else thinks or is likely to do.

The account *can* be characterized as a sort of two-systems view, of course. If our focus in making such a claim is on forms of mindreading that do, or do not, constitutively involve either the resources of working memory (above and beyond what is necessary to process the stimulus materials) or long-term memory, then that claim is trivial and uninteresting. For of course many cognitive systems can either operate from perceptual input alone or can do so when receiving input from, and contributing to the contents of, either working memory or long-term memory. The use of visual rotation to calculate what other people see is another matter, however. For this is a domain-specific executively-controlled mindreading strategy. And it is presumably one that children need to learn. In effect, they learn that when an asymmetric object is seen from different perspectives, one should visually rotate one's own image of it to figure out what the other person sees. But the resulting executively-controlled mindreading system is one that *encompasses* the automatic one, rather than being distinct from it. And no differences in conceptual resources are required. It seems less misleading to say that there is one mindreading system that can either operate together with executively-controlled resources (both domain-specific and domain-general) or not.[8]

Ideally this one-system account should be confirmed using fMRI. It predicts that we should see much of the same mindreading network active in implicit tasks that people have charted using explicit measures (Saxe et al. 2009). This includes medial prefrontal cortex, posterior cingulate, the superior temporal sulcus, and (especially) the temporo-parietal junction. Care will have to be taken, however, to insure that closely-matched tasks are used for purposes of subtraction. These should involve very similar

[7] In addition, the account sketched here comports nicely with the finding that people with autistic spectrum disorder do not display implicit false-belief understanding, although they can solve the very same tasks explicitly (Senju et al. 2009). It may be that such people either fail to encode beliefs automatically, or they do not have the standing goal of predicting agents' behavior, or both.

[8] Of course others, too, have claimed that mindreading depends partly on executive resources (e.g. Carlson et al. 2002, 2004). But this has been in connection with explicit (mostly verbal) mindreading tasks, which will make quite different demands on executive systems. As noted in Sect. 1, explicit tasks require participants to juggle and prioritize representations of the target agent's mental states, representations of the questioner's intent, and representations of the likely effect of their own replies on the mind of the questioner. None of this is true of implicit tasks. Nor are implicit tasks likely to require inhibition of a prepotent response.

sequences of observed activity, while differing from the target implicit-mindreading condition in not evoking spontaneous mindreading. Note, moreover, that on the one-system perspective presented here there should be little difference between implicit attributions of true and false belief. It is hardly surprising, then, that the one fMRI study to date of brain-activation in implicit mindreading found no activity in the temporo-parietal junction (Schneider et al. 2014b). For this chose to contrast the true-belief with the false-belief condition. Yet if the temporo-parietal junction is distinctively involved in encoding the thoughts of others, as many think (Young et al. 2010), then in *both* conditions the temporo-parietal junction will encode the agent's perceptual relation to a state of affairs and will use that as the basis for attributing a belief. The difference is just that in the false-belief condition the latter representation is not later updated.

Others have also proposed one-system accounts of mindreading, of course, while suggesting that the system needs to function together with executive resources in some circumstances (Leslie et al. 2004; Baillargeon et al. 2010, 2013). The present account is broadly consistent with these. But their focus was on explaining the differences in performance between infants and young children in implicit and explicit tasks respectively, where the executive demands are different from the cases we have been considering. (Notice, in particular, that while explicit tasks require executive inhibition of a reality-based response, nothing of the kind is needed in an implicit task of the standing-goal sort.) What I have tried to do here is show how a similar framework can be used to accommodate the recent data deriving from adults.

5 Conclusion

I have argued elsewhere that the recent infancy data do not provide evidence of two distinct systems for mindreading (Carruthers 2015). In the present paper I have defended a similar claim with regard to the adult data. The patterns of success and failure in the experiments that have been conducted with adults can be explained in terms of differing task requirements, specifically whether those tasks make significant executive demands of the participants. The data are best explained by an account that postulates just a single mindreading system which sometimes operates fully automatically, sometimes spontaneously in conjunction with long-term memory and the standing (unconscious) goal of anticipating people's behavior, and sometimes together with executively-controlled uses of working memory.

References

Apperly, I. (2011). *Mindreading*. New York: Psychology Press.

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

Apperly, I., Riggs, K., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*, 841–844.

Back, E., & Apperly, I. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, *115*, 54–70.

Baillargeon, R., He, Z., Setoh, P., Scott, R., Sloan, S., & Yang, D. (2013). False-belief understanding and why it matters: The social-acting hypothesis. In M. Banaji & S. Gelman (Eds.), *Navigating the social world*. Oxford: Oxford University Press.

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

Barrett, H., Broesch, T., Scott, R., He, Z., Baillargeon, R., Wu, D., et al. (2013). Early false-belief understanding in traditional non-western societies. *Proceedings of the Royal Society B (Biological Sciences)*, *280*, 1755.

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

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

Buttelmann, D., Over, H., Carpenter, M., & Tomasello, M. (2014). Eighteen-month-olds understand false beliefs in an unexpected-contents task. *Journal of Experimental Child Psychology*, *119*, 120–126.

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

Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.

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

Carlson, S., Moses, L., & Claxton, L. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental child Psychology*, *87*, 299–319.

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

Carruthers, P. (2015). Two systems for mindreading? *Review of Philosophy and Psychology*, *6*, 2–7.

Cohen, A., & German, T. (2009). Encoding of others' beliefs without overt instruction. *Cognition*, *111*, 356–363.

Csibra, G. (2007). Action mirroring and action understanding: An alternative account. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition: Attention and performance XXII*. Oxford: Oxford University Press.

De Bruin, L., & Newen, A. (2012). An association account of false belief understanding. *Cognition*, *123*, 240–259.

Dijksterhuis, A., & Aarts, H. (2010). Goals, attention, and (un)consciousness. *Annual Review of Psychology*, *61*, 467–490.

Ericsson, A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211–245.

Flavell, J. (1978). The development of knowledge about visual perception. In C. Keasey (Ed.), *The Nebraska symposium on motivation: Vol. 25. Social cognitive development*. Lincoln: University of Nebraska Press.

Gallagher, S., & Povinelli, D. (2012). Enactive and behavioral abstraction accounts of social understanding in chimpanzees, infants, and adults. *Review of Philosophy and Psychology*, *3*, 145–169.

Hanna, J., Tanenhaus, M., & Trueswell, J. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*, 43–61.

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

Huang, J., & Bargh, J. (2014). The selfish goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, *37*, 121–135.

Jacob, P. (2008). What do mirror neurons contribute to human social cognition? *Mind & Language*, *23*, 190–223.

Keysar, B., Barr, D., Balin, J., & Brauner, J. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–37.

Keysar, B., Lin, S., & Barr, D. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

Knudsen, B., & Liszkowski, U. (2012). 18-Month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, *17*, 672–691.

Kosslyn, S. (1994). *Image and brain*. Cambridge: MIT Press.

Kovács, Á., Kühn, S., Gergely, G., Csibra, G., & Brass, M. (2014). Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PLoS One*, e106558.

Kovács, Á., Téglás, E., & Endress, A. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*, 1830–1834.

Leslie, A., Friedman, O., & German, T. (2004). Core mechanisms in "theory of mind". *Trends in Cognitive Sciences*, *8*, 528–533.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551–556.

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.

Low, J., Drummond, W., Walmsley, A., & Wang, B. (2014). Representing how rabbits quack and competitors act: Limits on preschooler's efficient ability to track perspective. *Child Development*, *85*, 1519–1534.

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

Marien, H., Custers, R., Hassin, R., & Aarts, H. (2012). Unconscious goal activation and the hijacking of the executive function. *Journal of Personality and Social Psychology*, *103*, 399–415.

Onishi, K., & Baillargeon, R. (2005). Do 15-month-olds understand false beliefs? *Science*, *308*, 255–258.

Perner, J. (2010). Who took the cog out of cognitive science? Mentalism in an era of anti-cognitivism. In P. Frensch & R. Schwarzer (Eds.), *Cognition and neuropsychology: International perspectives on psychological science* (Vol. 1). New York: Psychology Press.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, *308*, 214–216.

Phillips, J., Ong, D., Surtees, A., Xin, Y., Williams, S., Saxe, R., et al. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *129*, 84–97.

Poulin-Dubois, D., & Chow, V. (2009). The effect of a looker's past reliability on infants' reasoning about beliefs. *Developmental Psychology*, *45*, 1576–1582.

Qureshi, A., Apperly, I., & Samson, D. (2010). Executive function is necessary for perspective-selection, not Level-1 visual perspective-calculation: Evidence from a dual-task study of adults. *Cognition*, *117*, 230–236.

Samson, D., Apperly, I., Braithwaite, J., Andrews, B., & Bodley Scott, S. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1255–1266.

Saxe, R., Whitfield-Gabrieli, S., Pelphrey, K., & Sholz, J. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Development*, *80*, 1197–1209.

Schneider, D., Bayliss, A., Becker, S., & Dux, P. (2012a). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*, 433–438.

Schneider, D., Lam, R., Bayliss, A., & Dux, P. (2012b). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, *23*, 842–847.

Schneider, D., Nott, Z., & Dux, P. (2014a). Task instructions and implicit theory of mind. *Cognition*, *133*, 43–47.

Schneider, D., Slaughter, V., Becker, S., & Dux, P. (2014b). Implicit false-belief processing in the human brain. *NeuroImage*, *101*, 268–275.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*, 883–885.

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

Scott, R., Baillargeon, R., Song, H., & Leslie, A. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, *61*, 366–395.

Scott, R., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. *Developmental Science*, *15*, 181–193.

Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., et al. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology*, *22*, 353–360.

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*, *22*, 878–880.

Song, H., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology*, *44*, 1789–1795.

Song, H., Onishi, K., Baillargeon, R., & Fisher, C. (2008). Can an actor's false belief be corrected by an appropriate communication? Psychological reasoning in 18.5-month-old infants. *Cognition*, *109*, 295–315.

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, *130*, 1–10.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*, 907–912.

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

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*, 580–586.

Surtees, A., & Apperly, I. (2012). Egocentrism and automatic perspective-taking in children and adults. *Child Development*, *83*, 452–460.

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, *129*, 426–438.

Surtees, A., Butterfill, S., & Apperly, I. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology*, *30*, 75–86.

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

van der Wel, R., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*, 128–133.

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

Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology*, *30*, 156–171.

Young, L., Dodell-Deder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, *48*, 2658–2664.