Belief Files in Theory of Mind Reasoning

Ágnes Melinda Kovács

Abstract Humans seem to readily track their conspecifics' mental states, such as their goals and beliefs from early infancy. However, the underlying cognitive architecture that enables such powerful abilities remains unclear. Here I will propose that a basic representational structure, *the belief file*, could provide the foundation for efficiently encoding, and updating information about, others' beliefs in online social interactions. I will discuss the representational possibilities offered by the belief file and the ways in which the repertoire of mental state reasoning is shaped by the characteristics of its constituents. A series of questions will be outlined concerning the representational skeleton of the belief file, sketching a possible structure that supports the rapid encoding and re-identification of belief related information (e.g., variables for the agent, as the belief holder and for the belief-content). After analyzing the possible limitations of the belief attribution system, I will examine some of its characteristics that might enable a flexibility that is often neglected. I will suggest that operations involving belief files are not impeded by the absence of precise first-person information regarding their contents. In fact, the system permits manipulations with "empty" belief files, allowing humans to ascribe beliefs to conspecifics based on little or no direct information regarding the content of the mental state. Such an analysis aims to advance our understanding of how spontaneous belief attribution may be performed, and to provide an insight into the possible mechanisms that allow humans to successfully navigate the social world.

1 Representations Underlying Theory of Mind Reasoning

Humans possess powerful capacities to encode and store various kinds of information, from different aspects of the environment to abstract conceptual knowledge. Such abilities are particularly impressive in young infants, who face a multitude of learning problems already in their first years or life. Amongst many other challenges, young learners have to acquire their native language, organize their knowledge about the environment into different categories and learn about the guiding principles of the physical and the social world.

Á. M. Kovács (✉)
Cognitive Development Centre, Central European University, Budapest, Hungary
e-mail: kovacsag@ceu.hu

One of the most amazing feats of humans is the ability to reason about other people's mental contents. We encode not only what we ourselves see, know and believe, but also what other people see, know and believe. Being able to take into account that people are guided by intentional mental states, such as goals and beliefs is usually termed as "theory of mind" (ToM) or "mindreading" ability. Such abilities allow us to predict and interpret others' behavior based on their mental states and ensure the success of online social interactions. While theory of mind abilities comprise reasoning about various mental states besides beliefs (e.g., goals, intentions, emotions), the term is most often used to refer to belief attribution (Premack and Woodruff 1978), a terminology also adopted here.

Remarkably, humans can attribute to another person any possible goal or belief they themselves can entertain, starting from the simplest ones, like assuming that two persons accidentally colliding in the morning rush did not notice each other, to more complex ones, such as religious, moral or political beliefs. Thus, ToM computations sometimes occur online, implicitly and spontaneously without much deliberation, and sometimes involve explicit, verbally expressed and often offline reasoning about mental states.

Recent debates regarding the nature of ToM abilities were inspired by developmental findings suggesting that implicit and explicit ToM processes may have different developmental paths. Research from the last 30 years has found that only children older than four give correct verbal answers regarding a character's behavior as a function of the character's false belief (Wellman et al. 2001). In contrast, more recent findings suggest that ToM processes are in place in early infancy when tested with implicit tasks (Onishi and Baillargeon 2005; Southgate et al. 2007; Surian et al. 2007; Kovács et al. 2010; Knudsen and Liszkowski 2012; Buttleman et al. 2009). To explain the findings that infants pass the implicit tasks while children fail the explicit ToM tasks before the age of four, researchers have proposed that implicit and explicit mindreading recruit different cognitive systems (Apperly and Butterfill 2009; Perner and Roessler 2012; Rakoczy 2012; de Bruin and Newen 2012). Such 'two-systems' approaches assume that spontaneous belief tracking relies on cognitive processes that are distinct from explicit ToM mechanisms. In most of these accounts, only the latter explicit ToM system is considered to operate on belief representations, whereas the implicit system is considered to involve associative processes (Perner and Ruffman 2005) or use simple relational information between agents, objects and locations (Butterfill and Apperly 2013; de Bruin and Newen 2012).

In contrast to such "two-system" approaches, here I will explore a different theoretical alternative according to which implicit ToM processes recruited in online social interactions, like explicit processes, require representing others' beliefs (Fodor 1992; Scholl and Leslie 2001; Sperber and Wilson 1995; Carruthers 2013). While there is a wide consensus that human adults can perform complex belief inferences and use sophisticated mental representations in an explicit manner, we know much less about the processes that are spontaneously recruited by implicit mentalizing, allowing even young children to navigate the social world. Hence, one of the biggest puzzles in ToM research is to understand what underlying processes make it possible to successfully track and update beliefs attributed to other people online.

In the present analysis, I will focus on the mechanisms and the representational apparatus that may allow us to readily compute other people's knowledge and beliefs. Specifically, I will outline the possibility that the precondition for spontaneous mindreading is a basic representational structure, called here the 'belief file' that

enables implicit ToM processes to store information about other agents' beliefs in a format supporting efficient encoding and updating. Before discussing in detail the structure and the functional properties of such belief files, I will first describe two examples that will allow us to examine the process of belief computation and to dissect the attributed belief representations into their possible constituents.

1.1 Example 1

1.1.1 The Problem of Time: Search for the 'Exact' Moment of Belief Attribution

Imagine a typical false belief situation, similar to the most frequently used standard ToM task, i.e., the Sally Anne task (Baron-Cohen et al. 1985; Wimmer and Perner 1983). You first see a character, Sally, putting her marble in a box and then leaving the scene. In her absence, Anne, the second protagonist, transfers the marble into a basket. Sally comes back, and when you are asked "Where will Sally look for her marble?" you will readily answer "In the box." While this answer reflects that you have correctly attributed to Sally the belief that the marble is in the box, it is unclear *when* exactly you have computed her belief. The moment at which the belief attribution takes place in such a situation can carry important information regarding the key features of the underlying mechanisms.

One can think of at least two theoretical possibilities regarding when Sally's belief could be inferred. The belief attribution could take place on demand, whenever it is necessary to predict or explain Sally's behavior, or it could be realized through the online monitoring of the world from Sally's point of view. According to the first alternative, which I will call *post hoc belief inference*, one computes Sally's belief only when this is required by the situation – for instance, when Sally returns to the scene and one wants to predict her behavior, or when one has to answer a question regarding her future behavior. Alternatively, one may track Sally's beliefs online, while the relevant events unfold, and sustain a representation of her belief despite the changes occurring in the real world. This alternative possibility can be termed *continuous belief tracking*.

The differences between post hoc inferences and continuous belief tracking pertain to *when, for what purpose*, and possibly in *how,* the belief is computed. In the case of post hoc belief inference, one is cued by a question, or by the reappearance of the agent, to infer her beliefs based on earlier events. In these cases one must recall the details encoded about the scene, analyze past events (i.e., the temporal order of the event sequences, the role different protagonists played in these events, whether they were present or absent during the most important scenes), and compute the agent's beliefs post hoc based on this information. In the continuous belief tracking scenario, in contrast, one spontaneously monitors what other people know and believe at all stages of the events. While continuous belief tracking refers to online and spontaneous computations of others' beliefs that happen under serious time pressure and it likely involves specific inferential processes, it is unclear whether these inferences are qualitatively different from post hoc belief inferences.

In trying to decide between these two alternatives, one could certainly argue for different views regarding parsimony and optimality in cognitive functioning. For instance, it could be more parsimonious to perform online belief tracking instead of post hoc belief inferences, as the latter requires "going back in time" and thinking of

past events. This would be particularly true for a developing cognitive system, as post hoc belief inferences presumably require more cognitive resources, and rely heavily on memory capacities that are not well developed in young children. On the other hand, one could argue that the most adaptive option would be to perform both kinds of computations (i.e., post hoc belief-inferences and continuous belief-tracking) flexibly, depending on the situation. While humans might be able to perform continuous belief tracking efficiently and effortlessly, there can be specific cases in which one is less likely to recruit such computations. It is unclear, for instance, whether in case multiple agents are present the cognitive system performs belief tracking for all the agents, or only for the situationally relevant ones. In the latter case, if an irrelevant agent (or an irrelevant event) turns out to be later relevant, one should be able to recall the earlier events, and based on these episodic memories infer whether an agent had some relevant experience allowing her to form a specific belief.

Experimental data from the last years has provided ample evidence that adults, children and even young infants can spontaneously predict a character's behavior based on her belief without being prompted to answer direct questions regarding her belief – that is, without involving a so-called 'elicited response' (He et al. 2012; Southgate et al. 2007; Senju et al. 2009; Rubio Fernandez 2013; Schneider et al. 2012; Helming et al. 2015). Such results exclude the possibility that one would *only* infer a character's belief when one is required to answer a question regarding her future behavior. Furthermore, studies have also found that it is not necessary that the agent return to the scene for the child to be able to compute the agent's beliefs; participants seem to encode the agent's belief online, while specific changes happen, and to sustain their representation of it even if the agent does not come back (Kovács et al. 2010). These findings suggest that one can form representations about other agents' mental states without having been prompted by the reappearance of those agents, or by questions regarding their beliefs. While there is little debate over whether human adults are able to infer other's beliefs deductively if they are required to do so, the evidence above suggests that human infants and adults can also spontaneously form representations about others' beliefs and sustain them in parallel with their own first person representations about the environment.

Certainly, the processes recruited by post hoc belief inferences and continuous belief tracking differ with respect to various dimensions. For instance, one is based on past happenings, while the other relies on online events; one is on demand, triggered by a question or by the need for a prediction, while the other is more spontaneous, possibly triggered by the mere presence of an agent. While it is an open question whether the two also differ in their underlying representations, in the present analysis I will mostly focus on the representational apparatus employed by spontaneous belief tracking. In the next example I will examine a real-life situation that may bring us closer to uncovering the structures that enable a fast and efficient tracking of others' beliefs, as well a continuous updating of representations about others' beliefs.

1.2 Example 2

1.2.1 The Problem of Speed: The Case of Multiple Consecutive Belief Updates

Imagine an everyday situation, where you are walking uphill with your colleague, engaged in an enthusiastic scientific debate. As you are approaching a grocery store,

you both notice that some shopping carts are blocking your way. Continuing the discussion, you notice that one cart is out of control and is starting to roll towards your colleague, who is positioned with her back to the cart and therefore does not see it, and who is in the middle of expounding her best argument. What would you do in such a situation? You would warn your friend about the oncoming shopping cart, or push her out of the way if you consider there to be immediate danger. However, if she happened to turn and look at the approaching shopping cart just as you were about to warn her, you would likely withhold or modulate your warning signal, and you would assume that she were now aware of the danger and preparing to perform the appropriate avoidance-behavior. However, if, after looking at the cart, she were to look back at you and simply continue her argumentation without moving out of the way, you may assume that she had been so caught up in the debate that she failed to notice the approaching cart. In this case, you would again update the belief you had attributed to her and return to your initial assumption that she were unaware of the cart, and warn her about the danger.

While we might encounter such situations, and even more complex ones, on a daily basis, note that updating an attributed belief multiple times involves a set of rather complex processes. In the span of a few seconds, we actually have to implement a large number of computations regarding what the other person believes, and about how she would behave on the basis of those beliefs. Furthermore, we use the outcome of these computations to flexibly adapt our own behavior as a function of the anticipated behavior of the other person. Such examples suggest that in online interactions, we readily update what other people see and believe (first we infer that our colleague did not see the cart, then we assume that she did, and finally we realize that she, in fact, is not aware of it), as well as what this means for their future behavior, and we rapidly modify our own behavior accordingly (prepare to warn her, suppress the warning, and finally provide a warning signal).

How is it possible to perform a whole chain of inferences and updates in such a limited time? It seems rather difficult to explain these multiple update processes in the post hoc belief inference framework. It is quite unlikely that we only infer our colleague's belief at the very last moment, right before she would get hit by the cart. Indeed, this would leave little time to update our representation of her belief multiple times. Presumably, in such situations we track others' beliefs spontaneously and update them in a continuous manner.

However, the issue of speed in online belief updating raises a further question regarding the underlying representational apparatus. Assume for a moment that, in the case of each update, we re-started the whole belief computation process from the very beginning. In the above example, this would mean initiating the whole process three times, i.e., starting with establishing that there is an agent who has or does not have visual access to the scene, and then going on to compute her belief content and its behavioral consequences. Presumably, such reiteration of belief inferences would not be fast enough to explain the efficiency of online interactions. However, this is what one should predict on the basis of a traditional view of ToM, according to which there is a unitary belief computation system, and each update would be a new belief attribution process.

On the other hand, there may be a more powerful alternative – specifically, that of a multi-component system that preserves those elements that need no updating (e.g., that

there is an agent who has a belief), and updates only some of the components (i.e., the belief-content in this case). In such a scenario, once the belief of the agent is computed, the belief representation may conform to a structure that permits the storage of the agent and the belief-content as separate constituents. Thus, the belief-content could be individually updated, while keeping the agent component constant. Such an update mechanism would result in a fast and efficient update, replacing just the critical component, while preserving the rest and ensuring the continuity of the process. Thus, decomposing attributed belief representations into their possible functional subcomponents could provide a fruitful framework for explaining how efficient belief updating can take place.

To provide a solution to the issues raised above, specifically to get closer in understanding how online belief tracking and updating may take place, I will be proposing here that representations of others' beliefs are stored in special representational structures –'belief files'[1]- that are sustained in parallel with one's own representations of the real world. The representational skeleton of these belief files allows for functionally separate subcomponents that can be individually updated, and for a format that supports the rapid encoding and identification of belief related information. Before moving on to discuss the characteristics of these representational structures, I will first address a more general issue, namely the reasons why ToM in general should not be regarded as an "all-or-nothing" ability.

1.3 Mentalization and the Underlying Cognitive Processes

The initiative to abandon the unitary view of ToM (where an individual either possess or does not possess ToM) can only be fruitful if it generates convincing proposals regarding the various processes that may subserve mentalizing abilities. Reflecting, for instance, on the scenarios described in Example 1 (targeting the exact time of belief formation) and Example 2 (targeting belief updating), it seems that we know very little about the possible cognitive mechanisms that underpin mental state reasoning. If an individual (e.g., a young child) fails on a standard false belief task, s/he might fail for a variety of reasons. Bloom and German (2000) suggested that different populations, such as atypically developing children and typically developing 3-year-olds, perform poorly on the standard false belief tasks due to different problems. They argue that typically developing children (but not children diagnosed with autism) might have difficulties, for instance, because of general task demands, because they don't have a grasp of beliefs being false, or because counterfactual reasoning might be too taxing for them. The list of the possible causes for failure, however, is far from exhaustive and can be extended to other possibilities. For instance, while children diagnosed with autism might have problems in conceiving of agents as having mental states and in forming belief representations, typically developing 3-year-olds or infants might have problems in sustaining the content of those belief representations that are in conflict with their own true beliefs. Therefore, if an individual shows difficulties in reasoning about others' mental states on such tasks, this may result from problems in conceiving that

---

[1] Here I will use the term belief files exclusively to refer to belief representations attributed to other people. In parallel, I will use the notion 'first-person (or regular) representations' to refer to one's own representations about the environment.

agents have mental states and in opening a belief file; from problems in computing the content of such a belief; from a failure in dealing with the conflict between different belief representations (one's own true belief and the attributed false belief), or from difficulties in making behavioral predictions. In consequence, identifying the possible component mechanisms of ToM processes and their underlying representations will not only bring us closer to understanding how efficient belief tracking and updating may take place, but also to understanding how specific processes unfold in development and whether specific populations have difficulties in some, but not other, ToM processes.

Thus, solving a typical false belief problem involves a larger set of cognitive processes, such as: i) opening a belief file; ii) computing the content of someone's belief; iii) binding belief representations to corresponding agents; iv) performing belief evaluations and pondering between true and false beliefs, and v) making implicit or explicit inferences upon false beliefs for behavioral predictions (see a detailed discussion in Kovács under revision). In the remainder of this paper, I will be focusing on a basic representational structure, the belief file, which may be a central constituent of these component processes and at the foundation of online belief tracking. Next, I will analyze the structural constituents and the possible functional roles of belief files.

2 Belief Files in ToM Reasoning

Belief files are theoretical constructs that aim to describe a core representational skeleton for online ToM reasoning – enabling us to efficiently track others' beliefs and to update our representations of those beliefs. How should we conceptualize such belief files? And how might the notion of a belief file contribute to solving current theoretical debates in the field? After a closer look at the possible structure of belief files, I will discuss evidence suggesting that implicit mentalizing seems to involve operations on the different constituents of these representational structures.

Successful ToM reasoning requires keeping track of other agents' mental states (e.g., their beliefs), which can be realized through opening a belief file that can then be filled with various contents. Belief files 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. In the event that only the content of the belief has to be updated, one will replace the content component, and will not have to re-initiate the whole belief computation process (i.e., one will not re-iterate the process of identifying that there is an agent who has a belief with a specific content, as discussed in Example 2 involving the shopping cart problem). In the following, I will briefly discuss the representational possibilities offered by belief files and the ways in which its repertoire is shaped by its constituents.

2.1 Storing and Individuating Belief Files

The aim in introducing the concept of a belief file is to advance our understanding of how others' beliefs might be encoded, and how those belief-attributions are updated. As we shall see, the concept of the belief file raises a series of new questions. For

instance, (i) it is unclear how we may store two related belief files, and (ii) once encoded, how a specific belief file can be re-identified.

Regarding the first question, efficient belief reasoning requires that different belief files be tracked separately, so one can perform specific inferences and updating processes upon them. However, it is not immediately clear what should be regarded as a single belief file, and in which cases more than one belief file is opened. To see why, consider the following pair of cases. The first case is a situation where two agents hold the same belief (e.g., both believe that an object is in location A), and it is likely that two belief files are opened. A separate belief file would be assigned to each agent, which can be updated individually in case only one of the agents has some relevant experience (e.g., sees that the object is transferred to location B), while the other agent does not see the event. However, the issue becomes more complicated in the second case in which we have, for instance, one single agent and two belief contents (e.g., Sally believes that Object 1 and Object 2 are in location A). Here, if a change happens only to one part of the belief content (only Object 2 is moved to location B) it is unclear whether this change would affect a single belief file (e.g., Belief 1: Sally believes that Object 1 is in location A, AND that Object 2 is in location B) or two belief files (e.g., Belief 1: Sally believes that Object 1 is in location A. Belief 2: Sally believes that Object 2 is in location B). One might argue that encoding such information in two belief files would allow faster updating in instances in which only one representation has to be updated. However, storing each content separately would lead to the accumulation of many belief files, possibly resulting in an encoding problem or a tracking problem. Alternatively, it may not only be belief-contents that can be updated separately from belief-holders (i.e., agents): it may be that different bits of information assembled within one belief content may also be individually changed. Using the example above, even if we encode in one belief file that "Sally believes that the cube is in location A, AND that the ball in location A" it is possible that one piece of information could be changed independently from the other.

When storing one or two belief files for the same agent, the cognitive system is faced with a 'compression dilemma', it is unclear what would be the most economic way of storage: the two-belief-file version, or the more compressed single-belief-file alternative. Given the lack of evidence supporting one or the other possibility, the actual level of fragmentation of the belief content remains hypothetical.

Turning now to the question of belief identification, once a belief file has been formed, the cognitive system should be able to re-identify it later in time, in order to use it for behavioral predictions or to update it. But how can we individuate belief files? While empirical studies are needed to elucidate what could serve as an index or 'address' for a specific belief file, I will describe one theoretical possibility. One way to understand the belief individuation problem would be to exploit the analogy between belief files and object files. Kahneman and Treisman (1984) introduced the concept of object file to explain processes underlying object individuation and object-based attention (see also Scholl and Leslie 1999; Scholl and Pylyshyn 1999; Pylyshyn 2001). According to this proposal, human infants and adults operate with mid-level representations of objects, via object files that are defined by their spatio-temporal characteristics. Each object has a unique spatio-temporal address, which can serve as an index or pointer to the object, and may or may not contain information about the features of the object. The object file system is precise and fast in tracking up to four

objects simultaneously, but it is unable to accommodate scenes beyond the limits of the indexes available. There is a variety of empirical and theoretical research supporting the idea of a continuity of the system from infancy to adulthood (Carey and Xu 2001; Leslie et al. 1998; Scholl and Leslie 1999).

While object files provide spatiotemporal indexes for object individuation, it is unclear whether belief files rely on an index system as well. Successful mentalization entails individuating, re-identifying and navigating among different belief files. An index system may be used to guide these processes. The most obvious candidate for individuating or indexing a belief file seems to be the agent to whom one attributes the belief. However, agents regularly uphold multiple beliefs. Hence, the agent alone does not seem sufficient to serve as an index with which specific belief files can be individuated. The next candidate that could serve as an index for belief files is the content of the belief. The same problems seem to apply here as well, as several agents may uphold the same belief content, and the content could be unspecified (e.g., Sally believes *something* is in the box, see part 2.2.). Thus, in this case the content would not constitute a very good index for identification.

Of course, for most belief files, both the content and the agent slots are well defined, such as representing that "Sally believes that the marble is in the box". Importantly, however, the representational format of the belief file allows that both the variable for the agent (e.g., "*Someone* believes that the marble is in the box") and for the content (e.g., "Sally believes that *something* is in the box") remain undefined. In consequence, it seems that neither the agent nor the content alone could be sufficient to individuate a specific belief file. While the minimal criterion for opening a belief file is the presence of an agent (even if not well-defined), belief individuation or belief indexing should rely on a relation between the belief-holder and the belief-content. Therefore, the index of a belief file, which would permit the individuation of an attributed belief, likely exploits a format that contains a conjunction of the [agent] variable and the [content] variable. It remains to be explored whether this possibility is already available early in development, or whether infants may use initially simpler strategies, such as performing, for instance, content-based belief individuation.

Finally, the analogy to object files has obvious limitations. While object files are characterized by a spatiotemporal pointer or index (which serves a deictic function like a finger pointing at an object, Pylyshyn 2001) belief files, in contrast, likely do not have such 'external' pointers (e.g., pointing to an agent). Instead, successful belief individuation might require an 'internal index' in the form of a conjunction between the agent and belief-content.

2.2 'Empty' Belief Files

In complex social situations one might not always have access to the content of another agent's mental state, or might not have enough cognitive resources for encoding it. Given the structure of the belief file, it may be possible to track another person's belief even if its exact content cannot be unequivocally determined when the belief is encoded. Hence, as discussed earlier, belief files can be defined as containing specific variables or placeholders that allow representation of an epistemic state of an agent A about a content X, where A and X can be replaced by various agents and a variety of contents, respectively, and can be defined to various degrees. In the following, I will refer to the case where the content is undefined as an "empty" belief file.

Traditionally, the cognitive systems involved in belief attribution have been considered to be prepared to deal with an infinite variety of contents. An interesting case for the present analysis is the situation when one attributes a belief to someone, and even updates this belief, without actually having computed the exact content of the belief. Consider now the following modified false belief scenario: Sally visually inspects the content of two identical opaque boxes, one of which has her favorite toy, but without the observer seeing what the boxes actually contain. Next, the typical intervention occurs: Sally leaves, and in her absence Anne, the second character, switches the boxes. Even if the exact content of the belief is undefined in this case (as the observer does not know where or what exactly the object is), the observer can still easily predict that Sally's behavior will now be governed by a false belief. If upon her return she choses one box, one can infer that (i) she believes the object to be there, and (ii) the object actually has to be in the other box, given that she has a false belief. Furthermore, if Sally then looks inside the chosen box, one will assume that – Sally's expectations not being fulfilled, as the object is not there – Sally must now update her belief. Thus, both opening a belief file and updating the belief-attribution can be performed without even knowing what the content of the belief actually is. When Sally returns to the scene one will attribute a false belief to her using an empty belief file, without knowing what she exactly believes about the location of the object.

Experimental evidence suggest that adults can readily perform such computations, and that they can attribute true and false beliefs to agents in similar situations while they themselves have no knowledge about the exact content of this belief (Apperly et al. 2004). In this study, Apperly et al. (2004) exposed participants to a scenario similar to the one describe above (used initially by Call and Tomasello 1999), where a character first looks into two opaque containers and then while she is away another character switches the positions of the containers. Then the first character comes back and points to one of the containers. Adult participants had the task to guess the real location of the object: the results showed that they could easily infer that the character had a false belief about the location of the object, and that they also correctly inferred the real location based on this false belief. Importantly, participants did not only assign a false belief to the character regarding the location of the object (without knowing where the object was hidden initially) – in fact, they also drew correct inferences about the actual state of affairs (where the object really was) based on the character's false belief. Note that the gold standard of how we think about belief attribution is the opposite of this. Specifically, we usually infer others' beliefs as a function of reality and what we believe about reality, not the other way around, inferring the actual state of affairs based on an attributed belief.

Thus, while we have already noted that humans are able to attribute to others any possible belief that they themselves can entertain the example above suggests that humans can also attribute to other people beliefs that they themselves *do not even entertain* (e.g., to recognize that some other person *knows* where the object is while we ourselves do not know where it is). I am not referring here to situations where one attributes a belief that one entertained at a previous time-point, such as "Sally believes the object is in the box", while one knows that the object was recently transferred from the box to the basket; or representations like "Sally believes in Santa Claus". Those are cases in which we have entertained the same belief content at an earlier moment in time or we entertain a proposition with an argument we know to be false. In contrast, the

present point concerns cases where we attribute belief representations to others regarding events we might be ignorant about, in the sense that we do not have a clear first person representation about them. For instance, in the above example when Sally returns to the room and chooses one of the boxes, one has not computed the representation that the object is in that location before attributing this belief to Sally. Furthermore, given that Sally didn't see the location switch and her belief must be false, one can conclude that the object is in fact in the other box. Remarkably, belief computation mechanisms not only allow us to infer what other people might believe about the world, but it seems that we can also use these belief inferences to learn new information about the state of affairs.

If belief inferences can result in acquiring new information, the issue of empty belief files turns out to be of crucial importance from a developmental point of view. Consequently, investigating whether preverbal infants can open empty belief files and fill in their contents later in time, not only could provide strong evidence for the debates regarding the origins of ToM, but could also be crucial in unveiling a social learning mechanism that exploits how our conspecifics see the world.

However, according to some recent proposals, young infants tested with implicit ToM paradigms do not perform belief computations but instead form three-way associations between the agent, the object and the location (Perner and Ruffman 2005), and recruit a so-called minimal theory of mind (Butterfill and Apperly 2013). It has been argued that while explicit ToM tasks involve a fully-fledged ToM that operates on belief representations, implicit tasks recruit a minimal ToM that relies on some simpler states that are not belief representations but make it possible to encode relations between the agent and its environment (Butterfill and Apperly 2013). If experimental evidence supports the possibility that infants are able to deal with empty belief files (where a relational encoding is not possible), such findings would suggest that core ToM components are present very early on in development, and that the basic representational structures used for belief tracking may be shared by infants and adults.

Consider, for instance, the earlier empty belief attribution situation, now adapted for infants: An infant is watching Agent 1 perform an invisible hiding event, putting an toy into one of two opaque boxes without the infant being able to see in which one. Then Agent 1 leaves the scene, and Agent 2 reveals the location of the toy, and she performs a second invisible hiding. Would young infants be able to infer the belief of Agent 1 when they themselves have no knowledge about the actual location of the object? Would they be able to compute at the moment of the first invisible hiding that the agent believes the object to be *somewhere* (open an empty belief file) and fill in its content later, when it is revealed by a different agent? In such cases, three way associations are excluded (infants never see the agent, object, and location together). One might wonder why should we expect infants to pass such a task, if children seem to succeed on the explicit version of this task only around the age of five (Call and Tomasello 1999). However, a growing body of evidence from the last 10 years suggests that standard ToM tasks can be adapted to measure infants' ToM abilities by transforming them in implicit tasks. Thus, it seems reasonable to ask whether infants would also succeed in an implicit version of a task that involves encoding empty belief files. In case future studies will suggest that infants can deal with such tasks, it would provide experimental support for the hypothesis that young infants, like adults, can operate with empty belief files, and treat belief contents and belief holders (agents) as separate constructs, and

hence use more sophisticated representations than simple relations among objects, locations and agents.

Note that I have used the term empty belief file in a lenient sense, specifically referring to cases that are characterized by under-defined belief contents. Empty belief files - in the form endorsed in this proposal – do not denote a case of 'representational vacuum', but instead refer to the capacity to exploit situations with limited or no information about the exact content of the belief one attributes to another agent, which can nevertheless assure an adequate inferential power. Thus, the term 'empty', in this context, does not indicate an empty set, but a referential uncertainty regarding identity, location or other properties of the possible belief contents. Of course, situational constraints may carry specific restrictions regarding what entities may apply as possible belief contents. For instance, in the case of the attribution "Sally believes *something* about the content of the box", the possible content of the belief file is most likely restricted to objects that fit inside the box. However, note that in theory one should be able to attribute any possible belief content regardless of whether it is reality congruent – such as, for instance: "Sally believes that there is an elephant/a witch in the box".

While it is an open research question whether we can talk of empty belief-contents in both a lenient sense (e.g., "Sally believes there is *something* in the box") and a strict sense (e.g., attributing that "Sally believes there is *nothing* in the box"), we most likely cannot talk of empty agent components in a strict sense. Belief-holders can be undefined, as in the example "*Someone* believes there is a marble in the box", but they cannot be completely open-ended. 'Someone' has to refer to an entity that is capable of having mental states, while belief-contents can refer to virtually anything, from real, likely objects (e.g., marbles), to real, unlikely objects (e.g., elephants), to impossible ones (e.g., witches) and maybe even to nothing. In a case where we find out that, in our belief that "Someone believes there is a marble in the box", the 'someone' slot turns out to be empty (no one, or not an agent), it is likely that we simply delete the belief file.[2] However, we probably do not delete the belief file in the case of "Sally believes something is in the box" when the 'something' slot turns out to be 'nothing'. I will return to such cases, and the possible limits of the system, in the following section, where the focus will be on how belief files may relate to other representations.

3 Belief files and First-Person Representations

Recently, the idea of 'mental files' has become a focus of attention (Fodor 2008; Recanati 2012), and motivated a series of proposals regarding the organization of mental representations in file-like representational structures (conceived like dossiers in a filing cabinet, or like encyclopedia entries that have a well-defined address and that store information regarding our knowledge of the world). Such proposals offer solutions regarding how different mental files may relate to each other, and about the nature of dependencies between attributed belief representations and first-person (or regular) representations most frequently concerning objects or events in our environment. The main question, then, pertains to the issue whether or not the representational structures

---

[2] This does not entail that we cannot entertain the proposition "No one believes there is a marble in the box", however, such computations might be outside the scope of belief tracking.

allowing mental state attribution are different from the ones used by our regular representational system dedicated to reflecting the world around us.

Although attributed and regular representations refer to the same reality, they often drive separate predictions regarding one's own behavior and the behavior of other agents, giving rise to a correspondence and detachment problem at the same time. According to one theoretical possibility, belief representations about objects attributed to other agents could be anchored to the actual objects, analogously to how we might anchor discourse referents to their external referents when the discourse is about actual objects (e.g., when pointing to a mouse and using the term 'mouse' or 'animal', depending on the context and the perspective one might take, Perner et al. 2007). Alternatively, attributed representations might be anchored to the corresponding first-person representation of the real object on the part of the person who makes the attribution (Leslie 1987; Recanati 2012[3]).

Leslie (1987) advanced an early proposal regarding the relation between these representations. The example that is most frequently used in Leslie's account is the understanding of pretense. He argues that, for instance, understanding the pretense action conveying that a banana is a telephone requires the establishment of detached representations. In these cases, the regular representation of 'the banana as a banana' must be decoupled from the representation of 'the banana as a telephone' in order to avoid representational confusion (that the banana is both a banana and a telephone). Leslie suggests that after the two representations are decoupled, they are linked via an informational relation. Specifically, he suggests: "Because decoupled expressions no longer automatically relate to the system of primary representation, they need to be specifically related to primary representations. Informational relations can be looked at as computational functions that perform this job, relating together agents, decoupled expressions and primary representations. One such informational relation is PRETEND, another might be THINK. PRETEND and THINK will differ in terms of the relationship they specify between agents, decoupled expressions and primary representations" (Leslie 1988, p28). As such, Leslie's account is a comprehensive proposal regarding general ToM reasoning, arguing that the representational structures involved in pretense are similar to the ones necessary for computing others' belief representations. Importantly, in this view, while both pretense and belief representations have to be decoupled from their primary representations, the link to their corresponding primary representations is realized through an informational relation.

Empty belief files, due to their nature, cannot exploit our first-person object representations. In consequence, proposals using a framework where the attributed belief must be linked to the real object (or to our own file about this object) will inevitably encounter a series of difficulties. For example, the cases discussed earlier, where ignorance is involved on the side of the attributer, might call for belief attribution with an empty/undefined content. This is true for situations in which, at the moment when one ascribes a belief or knowledge to another agent regarding an object, one lacks information about the object's identity and about its exact location. While it is difficult

---

[3] Recanati (2012) has argued that that there can be so-called 'unloaded' attributed files that are not linked to the subject's first-person files, such as for example in the case of attributing a belief to someone about Santa Claus in the case that the attributer does not believe in Santa Claus. However, one could argue that the observer should still have some kind of first-person mental file of Santa, based on which one can for instance answer the question whether Santa has a red coat – even apart from belief attribution.

to see how in this case the attributed belief file can be linked to the actual object, the limited information that is available (for instance, about the object's likely location) can still be exploited in the service of belief attribution.

Such examples draw attention to the limited role regular object representations may play in belief representation. The structure of belief files, by allowing for undefined variables, empowers us to cope with the uncertainty of these situations and to adaptively interpret others' action sequences based on these belief files. If we consider the invisible hiding scenario, even if the observer is ignorant about the object's exact location, he could possibly represent that some other agent (who had visual access to the hiding) believes the object to be in location "A or B". Later, if a location change happens during that other agent's absence, the observer will assume that the agent now holds a 'false' belief (without knowing its exact content), which will govern his behavior. If the agent searches in location A, one can at this point infer the content of the false belief and even figure out the real location of the object (not A, in this case, and thus B) by using a disjunctive syllogism.

In the following sections, I will outline some further cases derived from representational possibilities offered by belief files, with special focus on the relation between attributed and regular representations. In part 3.1. I will discuss a special case where a direct correspondence between a belief file about an object at a location and a first-person representation of that object (or the external object itself) is prevented because the external object ceases to exist (while another person still believes the object to be present). Then, in part 3.2. I will examine the possible differences between beliefs attributed to another agent about the presence of objects and beliefs about the absence of objects.

3.1 Belief Files and "Outdated" First-Person Representations

A further example pointing to the possibility that belief files can be sustained independently of first-person representations is, for instance, the situation where we observe an agent seeing an object being occluded and later the object dissolves without the agent seeing the dissolution. Although the object does not exist anymore, there should be no problem for us to sustain a belief file about the object attributed to the agent, thus representing that the agent still believes the object to be behind the occluder. This suggests that belief files can be readily sustained even when objects cease to exist, and thus our first-person representations of the object become outdated. Recent electrophysiological studies suggest that even 8-month-old infants seem to sustain such object representations attributed to others (Kampis et al. under review). In this study, the neural correlates of sustained object representation were measured in a task when an object was first occluded from an agent and then from the infant. The measured temporal gamma oscillations are considered to be a marker of sustained first-person object representations (Kaufman et al. 2003, 2005). Kampis et al. demonstrated that infants recruited the same brain mechanism when they themselves represented the continued existence of an occluded object and when they attributed such representations to another agent, and even in situations where after the occlusion from the agent the object dissolved. This suggests that young infants can sustain attributed belief representations independently from their first person representations about the reality.

In a similar vein, if an object does not exist anymore, it seems difficult to claim that the corresponding belief file is linked to the actual object.

The cases described above suggest that belief files might not necessarily be dependent on regular representations of reality or on actual objects in the environment. Of course, this is not to say that belief files are not usually formed in relation to the objects and events experienced in the environment or the corresponding first-person representations; the argument refers to the possibility that we can sustain belief files in the absence of first person object representations, and we can even update them, as discussed in the earlier sections of this paper. Such findings open the possibility that object files may not serve as a fundamental building block for the creation of belief files whose contents are about objects and their locations.

3.2 Belief Files and Absent Objects

Now consider a case where a belief attributed to another agent is not about the presence, but about the absence of objects. Representations of present and absent objects, however, may be treated differently within spontaneous belief tracking. In a study by Kovács et al. (2010), the belief of the participant that a ball is behind an occluder and the corresponding belief attributed to another agent (that the ball is behind the occluder), influenced participants' reaction times to a similar degree. This indicates that our own belief about an object and a belief we attribute to someone else with the same content rely on representational structures that allow for priming effects to occur. In contrast, the complementary belief attributed to another agent, specifically that the ball was *absent* (the ball exited the scene), had no effect on participants' reaction times.

There are two conjectures one can draw here. First, it is possible that the belief tracking system is more efficient in tracking others' beliefs about the presence of objects than beliefs about the absence of objects. Alternatively, belief tracking may take place both for objects that are present and objects that are absent, but the representational structures involved in the two attributions could have different contributions to the kind of effects elicited in this study (e.g., attributed beliefs about a ball being absent might not lead to priming). As a consequence, it seems reasonable to hypothesize that a set of constraints may apply to belief files formed during continuous belief tracking. Attributing beliefs about the absence of objects may pose representational demands that the implicit ToM system is not prepared to tackle. This possibility is also supported by recent neuroimaging data. Using an implicit ToM task, Kovács et al. (2014) found that only events involving false beliefs of an agent about the presence of an object, but not false beliefs about its absence elicited activation in the right temporo-parietal junction, an area that is regularly found to be selectively active in explicit ToM reasoning tasks as well (Saxe and Kanwisher 2003).

The differential effect of belief tracking involving present and absent objects (Kovács et al. 2014) and the early ability to encode another agent's representation of an object, despite first person experience about the ceased existence of the very same object (Kampis et al. under review), are phenomena that may shed light on complementary properties of the belief tracking system. The Kampis et al. study shows that infants possess powerful abilities to represent the content of someone else's false belief about an object being present, while knowing that the object does not exist any more. On the other hand, the Kovács et al. study points out some of the possible limits of this

system, by suggesting that it may have difficulties in representing another agent's belief that an object is absent, while the participants know it to be present.

There can be different reasons why the implicit ToM system may be more efficient in tracking false beliefs about the presence than false beliefs about the absence of objects. First, spontaneous mindreading might be functionally restricted to tracking behaviorally relevant mental state contents that could support real-time interactions. Indeed, only beliefs about the presence (and not the absence) of objects allow precise and efficient predictions regarding others' actions. For instance, knowing that Romeo falsely believes that Juliet is in the chapel would allow us to accurately predict Romeo's action (going to the chapel). However, attributing Romeo the false belief that Juliet is not in the chapel (although she is), does not enable us to exactly predict what Romeo would do next.

Alternatively, this phenomenon might not be a functional characteristic of the belief tracking system, but may instead stem from a genuine difference in how our cognitive system represents objects that are present versus objects that are absent. After all, while it seems rather straightforward how we might encode a scene where a ball goes behind an occluder, there can be a variety of ways in which we could represent that a ball goes out of sight. We may encode that the ball is away, it is in some other location, it is not in the scene, or we may not represent the ball at all, as it lost relevance when it went away. In this latter case, if the representation of the ball were deleted when it left the scene, the observation that we do not attribute to another person a representation, which we regularly do not encode for ourselves either would be little surprising. In any case, whether processing differently beliefs about the presence and about the absence of objects is an inherent characteristic of the belief tracking system, or a general representational feature of human cognition is a topic for future research.

4 Conclusions

In the present analysis I have focused on the belief file as a basic representational structure used in ToM, allowing on the one hand an efficient and continuous tracking of others' beliefs, and on the another hand also allowing for fast and multiple updates of representations of belief contents. Such updates are made possible by the skeleton of the belief file, where the content and the agent are conceived as separate slots. This makes it possible to rapidly update a belief content while keeping the other elements of the belief file constant (i.e., the agent). The centerpiece of this exploration was the concept of an empty belief file, i.e., a structure that allows for the content component to be tagged by a placeholder and, importantly, to fill in its content later in time. This flexibility may be crucial for a developing cognitive system that does not have the resources to encode all possible information at once. But even a mature system can make good use of such a feature, as there are many situations where the content of another agent's belief is not directly available. However, if we can encode that there is an agent who has a belief, although we don't know exactly what this belief might be, we can fill in this content information later, when it becomes available. Importantly, it is not clear how holding or updating a belief without knowing its content can be theoretically possible, unless one argues for a framework where attributing a belief and computing its content are separate components. Furthermore, an interesting feature

of empty belief files is that they can allow us to draw inferences not only about other's beliefs, but also about the actual state of affairs, based upon representations of others' beliefs.

Such flexibility of the cognitive system – that is, forming a belief representation without necessarily encoding its exact content – has also been proposed for the object domain. It has been argued that object files that stand for object representations, which do not necessarily contain information about the objects' exact features, enable a fast and efficient tracking of objects' motion. In a similar vein, belief files may serve an important role in social cognition, supporting efficient belief tracking and updating. The parallel of belief files with object files is only used as an illustrative analogy, as of course the two constructs may differ in a variety of ways. For instance, as discussed earlier, they have different indexes that allow identification, they have different underlying representational structures and they also likely have different encoding limits. The investigation of these differences may be a fruitful avenue for future research.

Opening a belief file through the continuous tracking and monitoring of others' beliefs might have various limitations, which might not apply to post hoc belief inferences. Consider, for instance, the example: "Nobody believes the cat is in the box". It is unlikely that the online belief tracking system is prepared to deal with such instances, given that they have no triggering properties (no agent). At a minimum, online belief formation requires a potential agent; if there is no agent, no belief file can be opened. Furthermore, while belief files seem to be useful when describing individual belief tracking, it is not clear how they can be invoked when there is not a single agent, but a group of agents. Hence, investigating the triggering conditions and the limits of forming and sustaining such belief files, would provide us with a better understanding of the processes that underlie ToM abilities in general. Future research should also be directed at unveiling whether empty belief files are equally available for human adults and young infants, and whether belief files use internal indexes and rely on representational structures such as the ones proposed here. Another issue for further consideration is whether representations underlying other kinds of mental states (such as goals or desires) could have a structure resembling that of belief files. While some of the questions raised in the present paper may be valid also for other mental states, these go beyond the scope of this proposal.

In sum, here I have argued that belief files constitute the basic organizational units of ToM reasoning that allow for efficient belief tracking and updating. The analysis presented here aimed to bring us closer to understanding how online belief attributions may be formed, encoded and updated; and how they relate to other kinds of mental representations. Such an exploration, at a minimum, suggests that implicit mentalizing, like explicit ToM processes, rely on belief representations, which I have been proposing to label 'belief files'. While belief files and explicit belief representations are formed and possibly accessed in different ways, the question as to whether they share the same representational format is an open one, and may be a target for examination in forthcoming studies. The initiative of specifying the various processes and representational structures underlying belief tracking may open up new directions for research – investigating, for instance, such representations in nonhuman animals, infants and populations with specific disorders, and will provide valuable new insights from finding which structures might or might not be present in specific populations.

References

Apperly, I.A., and S. Butterfill. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 116: 4.

Apperly, I.A., D. Samson, C. Chiavarino, and G.W. Humphreys. 2004. Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience* 16: 1773–1784.

Baron-Cohen, C., A.M. Leslie, and U. Frith. 1985. Does the autistic child have a 'theory of mind'? *Cognition* 21: 37–46.

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

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

Buttleman, D., M. Carpenter, and M. Tomasello. 2009. Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition* 112: 337–342.

Call, J., and M. Tomasello. 1999. A nonverbal false belief task: The performance of children and great apes. *Child Development* 70: 381–395.

Carey, S., and F. Xu. 2001. Infants' knowledge of objects: Beyond object files and object tracking. *Cognition* 80: 179–213.

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

de Bruin, L.C., and A. Newen. 2012. An association account of false-belief understanding. *Cognition* 123: 240–259.

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

Fodor, J.A. 2008. *LOT 2: The language of thought revisited*. Oxford: Oxford University Press.

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

Helming, K.A., B. Strickland and P. Jacob. 2015. Making sense of early false-belief understanding. *Trends in Cognitive Sciences*.

Kahneman, D., and A. Treisman. 1984. Changing views of attention and automaticity. In *Varieties of attention*, ed. R. Parasuraman and D. Davies, 29–61. New York: Academic.

Kampis, D., E.Parise, G. Csibra and A.M. Kovács (under review). EEG evidence for similar mechanisms to represent others' and own object representations in 8-month-old infants.

Kaufman, J., G. Csibra, and M.H. Johnson. 2003. Representing occluded objects in the human infant brain. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270(Suppl 2): S140–S143.

Kaufman, J., G. Csibra, and M.H. Johnson. 2005. Oscillatory activity in the infant brain reflects object maintenance. *Proceedings of the National Academy of Sciences of the United States of America* 102(42): 15271–15274.

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

Kovács, Á. M. (under review). Decomposing theory of mind: Belief files and further functional sub-components.

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

Kovács, Á.M., S. Kuehn, G. Gergely, G. Csibra, and M. Brass. 2014. Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PLoS ONE* 9: e106558. doi:10.1371/journal.pone.0106558.

Leslie, A.M. 1987. Pretense and representation: The origins of "theory of mind". *Psychological Review* 94: 412–426.

Leslie, A.M. 1988. Some implications of pretense for mechanisms underlying the child's theory of mind. In *Developing theories of mind*, ed. J. Astington, P. Harris, and D. Olson, 19–46. Cambridge: Cambridge University Press.

Leslie, A.M., F. Xu, P. Tremoulet, and B. Scholl. 1998. Indexing and the object concept: Developing 'what' and 'where' systems. *Trends in Cognitive Sciences* 2: 10–18.

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

Perner, J., and J. Roessler. 2012. From infants' to children's appreciation of belief. *Trends in Cognitive Sciences* 16: 519–525.

Perner, J., and T. Ruffman. 2005. Infant's insight into the mind: How deep? *Science* 308: 214–216.

Perner, J., B. Rendl, and A. Garnham. 2007. Objects of desire, thought, and reality: Problems of anchoring discourse referents in development. *Mind & Language* 22: 475–513.

Premack, D. G., & Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral & Brain Sciences* 1: 515–526. doi:10.1017/S0140525X00076512.

Pylyshyn, Z.W. 2001. Visual indexes, preconceptual objects, and situated vision. *Cognition* 80: 127–158.

Rakoczy, H. 2012. Do infants have a theory of mind? *British Journal of Developmental Psychology* 30: 59–74.

Recanati, F. 2012. *Mental files*. Oxford: Oxford University Press.

Rubio Fernandez, P. 2013. Perspective tracking in progress: Do not disturb. *Cognition* 129: 264–272.

Saxe, R., and N. Kanwisher. 2003. People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage* 19: 1835–1842.

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

Scholl, B.J., and A. Leslie. 1999. Explaining the infant's object concept: Beyond the perception/cognition dichotomy. In *What is cognitive science?* ed. E. Lepore and Z. Pylyshyn, 26–73. Oxford: Blackwell.

Scholl, B.J., and A.M. Leslie. 2001. Minds, modules, and metaanalysis. *Child Development* 72: 696–701.

Scholl, B.J., and Z.W. Pylyshyn. 1999. Tracking multiple objects through occlusion: Clues to visual objecthood. *Cognitive Psychology* 38: 259–280.

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

Southgate, V., A. Senju, and G. Csibra. 2007. Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science* 18: 587–192.

Sperber, D., and D. Wilson. 1995. *Relevance: Communication and cognition*. Oxford: Blackwell.

Surian, L., S. Caldi, and D. Sperber. 2007. Attribution of beliefs to 13-month-old infants. *Psychological Science* 18: 580–586.

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

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