The representational flexibility of spontaneous theory of mind in human adults

Dóra Fogd

Central European University

Department of Cognitive Science

In partial fulfilment of the requirements for the degree of Doctor of Philosophy in Cognitive Science

Primary supervisor: Ágnes Melinda Kovács

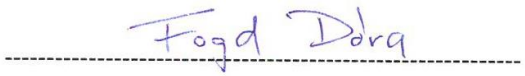
Secondary supervisors: Natalie Sebanz and Ernő Téglás

Vienna, February 2023

Declaration of Authorship

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no materials previously published or written by another person, or which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgment is made in the form of bibliographical reference.

I also declare that the intellectual content of this thesis is the product of my own work, and the claims here reflect my own thinking. However, as I believe scientific research is inherently a social activity, and all work receives support from others in theoretical, methodological, or stylistic matters; I will present this work in first- person plural voice.



Dóra Anna Fogd

02/27/2023

Abstract

Successful social interactions require correct interpretations and predictions of others' actions. It has been widely accepted that to be able to do so people rely on their capacity to represent other agents' mental states and take into account that those may differ from their own, referred to as theory of mind (ToM) in the literature. Research from the last fifteen years indicates that adults as well as infants may represent the content of others' false beliefs and visual perspectives spontaneously, however little is known about how other types of ToM computations take place in adults. Here we ask whether, contrary to some theoretical proposals, ToM can be flexible and efficient at the same time, in a sense that it can track various contents and allow for the complex manipulation of attributed beliefs, spontaneously.

The present thesis investigates three ToM processes that may play an important role in the smooth adaptation to others' behaviour in a number of everyday social interactions, yet have been so far unexplored: 1) the updating of other agents' mental states on the basis of the behaviour they demonstrate in a situation (Chapter 2), 2) the encoding of the hypotheses (alternatives) they entertain (Chapter 3 and 4) and 3) the representation of the conclusions they may draw from the beliefs they hold (Chapter 4). Specifically, in each study, we asked whether these computations take place spontaneously, even when people are not required to perform those.

Using anticipatory looking and behavioural measures, in Study 1 we found evidence that adults spontaneously update a previously attributed mental state of another agent and revise their expectations regarding the agent's future behaviour if they observe the agent repeatedly performing actions incompatible with their original assumptions about the beliefs she/he holds. Regarding our second research question, in Study 2 we did not find strong evidence for the spontaneous encoding of the alternatives another agent may represent in a situation with a change detection paradigm. However, we did find convincing evidence for the spontaneous representation of such contents in Study 3, in a series of online experiments using a different task, where participants had to estimate the likelihood of certain events, from first- and third-person perspective. Importantly, the results of Study 3 also suggest that human adults' spontaneous ToM abilities go beyond monitoring what others saw, and consequently know, and extend to the representation of the conclusions other agents may deductively draw, from the beliefs they hold. In this set of studies, we found that adults spontaneously tracked others' logical inferences both about object identity and location, even if those involved multiple steps, if the cognitive demands of the task were relatively low.

Taken together these results imply that human adults are endowed with rich spontaneous ToM abilities that allow them to function *both efficiently and flexibly* in the social world. Nevertheless, there are large individual differences regarding whether they actually perform these computations spontaneously, which highlights the importance of investigating individual heterogeneity in the use of ToM abilities and its underlying factors.

Acknowledgements

First, I would like to thank all the guidance and encouragement I received from my primary supervisor, dr. Ágnes Melinda Kovács throughout the years. Without her support I would not have been able to finish my studies. She devoted an immense amount of time to the studies presented in the current thesis and she was always there for me at times of failures and personal difficulties. I am also grateful to dr. Natalie Sebanz, especially for her valuable comments on the work presented in Chapter 2, and her always positive attitude. I am indebted to dr. Ernő Téglás, for his contribution to the work included in Chapter 3 and Chapter 4 of this thesis. His theoretical and methodological rigor serves as a true example for me. I am grateful for the opportunity that I could work with him.

I would like to thank the past and current members of the Cognitive Development Center for the thoughtful comments I received from them throughout the years. And of course, for their willingness to participate in the numerous pilot studies I ran to develop the final versions of the design and the stimuli. I owe special thanks to Tibor Tauzin and Oana Stanciu, for the help they provided in making sense of some of my messy data as well as for their constant encouragement. I am also grateful to Eszter Szabó, Eszter Körtvélyesi, Atsuko Tominaga, Nazli Altinok and Gabi Felhősi for the invaluable emotional support I received from them in the last few years. I am glad that I can call them friends.

I would also like to thank all those volunteers who allocated their time to participate in my (always too long) experiments. Part of the work presented in the current thesis was funded by grant Ágnes Melinda Kovács and Ernő Téglás received from the European Research Council. I am grateful for the funding.

I am thankful to Réka Finta for her kindness and for all the administrative help she provided during my studies. She is a true wizard who can solve all problems in no time. I could not wish for a better departmental coordinator.

Last, but not least I would like to thank my parents the incredible amount of support they provided for me, from the moment I stepped in their life at age four until their death. Despite both of them passed away during the years of PhD, their love remained with me and helped me through everything. I hope I managed to serve their faith in me.

Table of Contents

Chapter 1: Theoretical framework

1.1 Introduction

Humans are ultrasocial beings: they spend most of their time engaging in various cooperative interactions with the members of their group. From playing basketball through navigating in a crowded street to trying to convince others in a debate, they face a multitude of situations where the ultimate success depends on the appropriate interpretation of others' behaviour (why they move or say what they do) and the correct anticipation of their future actions, at various moments, during the entire course of the interaction.

Although in some cases it is possible to tell what the other will do next and come up with an explanation for the observed actions based on simple rules, derived from statistical regularities, or social scripts about how events usually unfold, it has been widely accepted that humans explain and predict other agents' actions in terms of unobservable mental states. They attribute goals, beliefs and desires to others, expecting them to act in line with those. The capacity to represent other agents' mental states and take those into account when interpreting and predicting their actions, has been termed *theory of mind* (ToM) (Premack & Woodruff, 1978) or mentalizing and has been considered to be one of the most important higher-order cognitive abilities of the human species, given the role it arguably plays in successful coordination, communication and the acquisition of the building blocks of culture (Tomasello et al., 2005; Herrmann et al., 2007). Importantly, ToM entails understanding that beliefs are mere representations of reality, therefore they might not reflect the true state of affairs and they may not coincide with the observer's point of view. Crucially, however, in situations where the agent holds a true belief, it is impossible to tell apart whether correct prediction of his/her behaviour relied on the content of the attributed belief or the observer's own representation of the state of the world. This has led researchers, striving to find a proof for the 'representational' understanding of the mind, to narrow down the study of ToM abilities to the investigation of people's capacity to *attribute false beliefs* to other agents. Specifically, for decades, it led them to test people's performance on tasks that require participants to predict how an agent, who is mistaken about an object's actual location or identity, will act.

Theory of mind has been the target of extensive research in the past 40 years. Numerous studies have investigated the ontogenetic (Wellman, 2018) and phylogenetic origins of false belief understanding (Martin & Santos, 2016), as well as its neural basis (Frith & Frith, 2006) and putative deficits in various atypical populations, such as people living with autism (Baron-Cohen, 2000), with somewhat less attention directed at the question how the attribution of such beliefs takes place in typical adults. Early studies relied almost exclusively on verbal responses provided in some version of the standard false belief task for direct questions about the other agent's future behaviour and belief. These studies

pointed out serious limitations of young children's (Wellman et al., 2001), and atypical populations' (Baron-Cohen et al., 1985) capacity to attribute false beliefs to and to understand the incongruent visual perspective of other agents, as well as in adults' ability to use the attributed content for predicting others behaviour (Keysar et al., 2003). On the contrary, more recent studies, using various nonverbal measures to capture the 'spontaneous' operation of ToM, demonstrated a remarkable sensitivity to the content of another agent's (false) belief even in young infants (Scott & Baillargeon, 2017) and nonhuman primates (Krupenye & Call, 2019). In addition, they provided a number of evidence that belief attribution (or more generally perspective-taking) occurs in a fast and efficient manner in adults (Schneider et al., 2017). These findings have generated a heated debate regarding the representational underpinnings of the ToM abilities manifested in the 'nonverbal ToM' tasks. This, in turn, has led to a new line of research in an attempt to establish whether results reflect the functioning of a full-blown theory of mind, or a minimalist version of ToM that allows behaviour prediction in some cases, but does not entail understanding the representational nature of mental states (Butterfill & Apperly, 2013), or in fact reflect the operation of non-mentalistic processes (Heyes, 2014a). What is more important for the purpose of the present thesis, it shifted attention to the question, *what is the nature of the ToM processes in general*, i.e. to what extent and in what sense can the attribution of epistemic mental states, such as knowledge or belief be considered 'automatic' in humans.

Most of the authors who claim that the above-mentioned findings (from studies using nonverbal tasks with adults and infants) reflect the operation of a mature ToM mechanism consider mentalizing an automatic or quasi-automatic process. They argue that it is specifically the rapid, involuntary nature of mental state attribution that makes it possible for humans to successfully interact with others (e.g. Carruthers, 2017; Kovács, 2016). This is what allows people to quickly formulate appropriate expectations regarding others' future actions and readily adapt their own behaviour to the outcome of the predictions. While the ability to track other agents' perceptual (or more broadly: informational) access and to compute what they see or believe to be true at a given moment on the basis of this information, in a quasi-automatic manner, may provide the necessary basis for the smooth unfolding of social interactions, it might not be sufficient to ensure swift adaptation in a number of situations, that frequently ensue in daily life. Such situations arise when some aspect of the environment (that likely has an impact on the other's mental state content) or the other's behaviour rapidly changes, when the agent who is about to act is uncertain about the current state of affairs or when others' actions are based on the inferences they have drawn from the beliefs they hold rather than on the directly available information. These require additional ToM abilities that have been so far largely unexplored: the capacity to (1) update the previously attributed mental state content and revise one's expectations in line with the outcome of the process (after realizing that there is a need to do so), to

(2) represent the alternatives or the 'hypotheses' the other agent does and to (3) track what conclusions others may or may not draw from the beliefs they hold. Not only should humans possess the ability to perform these computations, but they should also be able to recruit these abilities spontaneously in the relevant situations, to be able to adapt smoothly to the observed changes and/or (re)act fast enough when it becomes necessary in the future.

To this end the present thesis investigates the question whether human adults indeed perform the above-mentioned computations (update others' mental states, encode the content of their hypothesis space, and represent the conclusions they can draw) spontaneously, in a similar manner as they seem to attribute false beliefs to other agents. More broadly, it explores the question how theory of mind capacities contributes to the remarkable flexibility humans demonstrate in online social interactions. In the following sections of Chapter 1 we will provide an in-depth discussion of the theoretical debate on the nature of theory of mind, together with a review of the related empirical findings. Then we will identify some of the open questions in the field and finally present the aims of the present thesis.

1.2 From explicit to implicit theory of mind: findings from children and adults

1.2.1 False belief understanding in children

For more than 20 years the primary tool for investigating theory of mind abilities was the standard false belief task, in particular, the unexpected change-of-location task, and the probe for understanding beliefs was whether participants provided correct verbal responses in this task. In the most widespread version of the task, the so-called Sally-Ann task (Baron-Cohen et al., 1985), participants are presented with a scenario in which the protagonist's (Sally's) toy is transferred from one container into another by her friend (Ann) in her absence. Upon Sally's return, participants are asked to predict where Sally will look for the toy and provide justification for their answer.

Based on findings from various unexpected change-of-location tasks (and some other versions of the false belief task), it was proposed that the ability to represent mental states develops around the age of four (Wellman et al., 2001). Children younger than this age show systematic failures on the task, predicting that the protagonist will look for the toy where it really is, i.e. base their prediction on the actual state of affairs and/or their own knowledge instead of the agent's false belief - a phenomenon usually termed as 'pull-of-the-real' (see: Carpenter et al., 2002). Some authors argued that the basic representational architecture for attributing false beliefs is present from birth, in the form of an innate

module ('Theory of Mind Module'), and the reason why young children do not pass the task is that executive processes ('Selection Processor'), that could help them to overcome the default 'reality bias', are not mature yet (Leslie et al., 2004). Although some results indicate that the cognitive load imposed by the standard task indeed plays a role in three-year-olds' performance (Carpenter et al., 2002; Rubio-Fernández & Geurts, 2013; Setoh et al., 2016), providing evidence for this account, for decades, the most widely accepted explanation of the findings was that children's 'theory' of the mind undergoes a radical conceptual change around age four (Gopnik & Wellman, 1992; Rakoczy, 2017). Passing the false belief task was considered to mark the emergence of the ability to represent *beliefs as representations*, more specifically *as attitudes towards a proposition p* (Perner, 1988), and, to understand the critical features of representational mental states – that they can differ from one's own, they can misrepresent reality and that they represent reality under some aspect (for a detailed discussion see e.g.: Rakoczy, 2017). Results showing that children start to pass the Level-2 visual perspective-taking task (i.e. understand *how* exactly an object or a scene looks from a different point of view) and the so-called appearance-reality tasks around the same age (Flavell et al., 1981; Moll & Meltzoff, 2011), both of which requires understanding that one thing can be conceptualized in different ways (that a sponge can look like a rock or that a person sitting on the other side of a table sees things upside down), lent support to this view.

Given the focus of the early research, few studies have investigated how ToM develops beyond the preschool years. The ones which did so focused on the question *when* different milestones, such as the understanding of higher-order false beliefs or the recognition of faux pas, are achieved. These studies revealed that ToM abilities continue to develop after early childhood (Sullivan et al., 1994; Baron-Cohen et al., 1999), until late adolescence, with some tasks, e.g. higher-order false belief tasks that involve several levels of recursion, posing serious challenges even for adolescents (Valle et al., 2015).

1.2.2 The process of belief attribution in adults

Despite the vast amount of studies that have been conducted in the field, for decades, the question *how* belief attribution actually takes place, i.e. whether beliefs are ascribed automatically by older children and human adults (whenever an agent is present) or only when they are prompted to do so or when it is necessary for a task, remained unexplored. The first studies tried to approach this question by investigating whether the processing of unexpected statements about an agent's (false) beliefs incurs an extra cost compared to the processing of unexpected statements about reality. Using

this methodology, Apperly and colleagues (2006) found that adult participants, who were watching true and false belief scenarios, with the instruction to indicate the location of the hidden object at the end (thus had no motive to represent the belief of the agent), were slower to respond to incidental probes about the protagonist's belief than matched probes about the location of the object. Based on these results, the authors concluded that belief attribution is most likely non-automatic. Others questioned the interpretation of these findings, pointing out that, in case of belief probes, a substantial time elapsed between the probe and the timepoint the agent indicated her belief about the location of the object (by placing the cue on the respective box before she left), which might have led to the decay of the information from working memory, even if the agent's belief was originally encoded. Indeed, when the delay between these two events was shortened by changing the event sequence such that the agent gave the cue of her belief after the boxes were switched, Cohen and German (2009) found that responses to belief probes were even faster than responses to probes about reality, and just as fast as the responses participants provided to the belief probes when they were instructed to track the agent's belief content. Results indicating a similar processing cost for true beliefs, in a picture-sequence version of the original task, i.e. in the absence of timing differences between the reality and belief probes (Back & Apperly, 2010), have, however, challenged this account. The authors raised the possibility that in Cohen & German's (2009) version of the task, unlike in the original study, participants might have been (accidentally) prompted to compute the agent's false belief, given that the agent cued the wrong box after the switch. This act triggered the otherwise non-automatic mentalizing process to find an explanation for the 'mistaken action' of the agent.

These findings, together with others indicating a serious impairment of adults' performance on complex ToM tasks when a secondary executive task is administered (see e.g. McKinnon & Moskovitsch, 2007), led to the general conclusion that belief attribution relies on cognitively demanding computations. Therefore, it was proposed that belief attribution takes place only when prompted, either by instruction or by the task context that motivates adults to interpret the behaviour of the other (Low et al., 2016). In line with this, a number of studies has shown that adults are slower and tend to make more errors when they have to take into account the speaker's visual perspective, knowledge or false belief to disambiguate the referent, on the so-called 'Director Task' (Epley et al., 2004; Keysar et al., 2000; Keysar et al., 2003): they look at first and/or manipulate the object that matches the description from their own and not from the interlocutor's perspective. These results were taken as further, albeit indirect[1] evidence that mental state attribution is indeed a non-automatic process.

[1] Note that these task measure how people *use* ToM in language comprehension, more specifically how they take the other's visual perspective into account to disambiguate a referent, which requires participants to make further inferences after the other agent's mental state content (what the agent can and cannot see) has been

1.2.3 The paradigmatic shift in the study of ToM

The view that belief attribution is a slow, effortful, deliberate process with the ability emerging rather late in development, has been seriously challenged in the last 15 years. First, a number of developmental studies found that if 'indirect' measures are used, such as looking time, spontaneous helping or anticipatory gaze, and nonverbal versions of the standard false belief task, even toddlers and infants demonstrate a sensitivity to other agents' false beliefs (for reviews see: Barone et al., 2019; Scott & Baillargeon, 2017). For example, early studies using the violation-of-expectation paradigm found that 15-month (and even 13-month) old infants look longer if the protagonist's action was inconsistent with her false belief, i.e. searched at the correct location following an invisible object-transfer (Onishi & Baillargeon, 2005; Surian et al., 2007). Others, using a different task, found that infants look longer even at an expected outcome (the object not being present) if it contradicts the false belief of an agent, at 7 months of age (Kovács et al., 2010). Another line of studies demonstrated that 18–24-month-old toddlers spontaneously help the agent to achieve his/her goal, they could infer (only) on the basis of the agent's false belief, e.g. open the intended box or inform the other about the desired object's correct location to prevent mistake (Buttelmann et al., 2009; Knudsen & Liszkowski, 2012a). Finally, multiple studies found that toddlers tend to look more towards the location corresponding to the protagonist's false belief, than towards the location they themselves last saw the object, right before the protagonist would act, which was taken as evidence that they expect the agents to search where they believe the object to be (for the original finding see: Southgate et al., 2007). It is important to note that the replicability of some of these measures (in demonstrating false belief understanding) was seriously questioned lately (see e.g. Kulke, Reiss et al., 2018), with some authors pointing out that, for instance, anticipatory looking tasks – at least those involving object-transfer and measuring looks before the agent would reach - fail to even elicit spontaneous action prediction, in many cases, even when the agent has a true belief (Baillargeon et al., 2018; Kampis et al, 2021). Nevertheless, for a long time, results from these tasks, together with the rich body of consistent and convergent findings from other paradigms, were considered as strong evidence for the presence of *'implicit' ToM* abilities in infancy, with some authors referring merely to the mode of measurement, others to the nature of the representational architecture with this terminology.

With respect to adults, a number of studies have found that the mere presence of another agent holding a divergent belief or a conflicting visual perspective modulates adults' performance on various tasks, even if the agent's mental state is irrelevant for the task to be performed (Buttelmann &

computed. This leaves open the possibility that mistakes and slower responses reflect difficulties with these additional steps and not with the mental state attribution per se.

Buttelmann, 2017; Kampis & Southgate, 2020; Schneider et al., 2017). In one of the first studies demonstrating this effect, Samson and colleagues (2010) presented adult participants with pictures displaying a room with varying number of discs on the walls and a human avatar, who, by virtue of her orientation, perceived either the same or a fewer number of discs as the participant. Participants' task was to judge the number of the discs either from their own or the avatar's perspective, as fast as possible. The authors found that, in cases where the two perspectives conflicted, participants were slower and more prone to errors not only on trials where they had to judge the avatar's perspective (demonstrating the well-known egocentric bias) but also when they performed self-perspective judgements. This effect, which has become known as 'altercentric interference' in the literature, also emerged when participants only had to judge their own perspective and never that of the avatar (Samson et al., 2010, Experiment 3), and was demonstrated later in a number of other visual perspective-taking, and, recently, even in a nonverbal false belief task, with a trial structure and design very similar to that of the original task (Meert et al., 2017).

Using an object detection paradigm, Kovács et al. (2010) found an opposite, 'priming' effect of a task-irrelevant agent's belief content on the performance of adults: participants were faster to detect the presence of a ball they themselves saw leaving if the agent believed the ball to be present (behind an occluder), compared to the situation when both they themselves and the agent expected the ball to be absent. These findings were replicated in a number of studies, using the same paradigm (El Kaddouri et al., 2020; Nijhof et al., 2016; Nijhof et al., 2018). A similar 'facilitatory effect' of the other's visual perspective content was demonstrated recently in a perceptual decision-making task, by Ward and colleagues (2019), suggesting that this effect is not confined to one type of mental state content and one specific paradigm. Specifically, the authors found that participants were faster to judge the form of rotated letters in the presence of a task-irrelevant person, if those appeared in a close-to-canonical orientation to the other, i.e. when judgements were easier from the other person's view, indicating that they represented the content of the other's visual perspective automatically, in a quasi-perceptual form.

Besides affecting performance on speeded behavioural tasks, the content of another agent's belief was also found to influence participants' movement trajectories, in a mouse-tracking (van der Wel et al., 2014) as well as their looking behaviour in adult versions of the infant anticipatory looking tasks (see e.g. Schneider, Bayliss, et al., 2012; Schneider et al., 2014), generating a bias towards the location consistent with the agent's false belief. Importantly, these effects emerged even if participants were not instructed to track the agent's belief and were completely unaware of doing so, suggesting that belief computation is an unintentional and to some extent, unconscious process.

Altogether, these findings led many authors to conclude that, in line with our everyday intuition, in human adults, mental state attribution takes place in a fast, efficient, possibly automatic or quasi-

automatic manner. Results showing that performing a concurrent inhibitory control task does not eliminate the 'altercentric interference effect' on the 'dot-perspective-taking task' (Qureshi et al., 2010), lent further support for this interpretation of the findings (though for an opposite finding see: Schneider, Lam, et al., 2012) and provided evidence that, unlike how it was previously assumed, the calculation of other agents' (visual) perspective takes place effortlessly in humans.

1.3 The nature of theory of mind: theoretical debates

While the early years of ToM research were dominated by the debate on what young children's failure on the standard false-belief task reflects, the conflicting results presented in the previous section shifted attention to a new theoretical question: whether the mechanism that subserves the often fairly complex explicit belief inferences of adults and enables children to pass verbal ToM tasks can be the same as the one that underlies infants' and adults' performance in the nonverbal ToM paradigms and in everyday interactions. In other words, whether efficiency and flexibility can co-exist in the operation of theory of mind.

1.3.1 Nonmentalistic accounts

The interpretation of findings from the so-called 'nonverbal' ToM tasks generated considerable debate in the last 10 years, with some authors arguing that they do not reflect mentalizing at all. For example, Ruffman and Perner (2005) claim that infants succeed on such tasks (specifically on those involving object-transfer) by forming associations between agents, objects and locations or by applying certain 'behaviour rules' (such as 'people search for objects where they last saw them'), derived from statistical regularities (for detailed discussion see: and Ruffman, 2014). Others argue that most of the results, both from infant and adult studies, can be explained by low-level, domain-general processes. In particular, Heyes (2014b) claims that infants' looking behaviour results from the 'novelty' of the test events compared to those encoded earlier, caused by the perceptual salience of the test stimuli or the disruption of memory processes that make infants forget previous events. In a detailed review, she provides an alternative account for each of the early findings, with the exact cause why the specific test event would count as 'novel' for the infant differing in each and every explanation offered. The interference and priming effects obtained in adult studies are argued to reflect similar attentional and

memory processes: attentional orienting, elicited by the directional cues of the agent in Samson and colleagues' (2010) study and retroactive interference caused by a perceptually salient event, the reappearance of the agent in the critical condition of Kovács and colleagues' (2010) experiments (Heyes, 2014a). While the latter account seems to be unlikely, based on the Kovács and colleagues' (2010) own results from a control experiment, matching the conditions with respect to when the ball is last seen by the participants (discussed in their Supplementary Materials), the former received some support from experiments showing that arrows elicit a similar interference effect as human avatars (Santiesteban et al., 2014). Finally, some authors questioned the results of specific studies, pointing out that they may reflect confounds. For instance, Phillips and colleagues (2015) claim that the priming effect observed in Kovács and colleagues' study results merely from the timing of the attention check involved by the authors to ensure that participants pay attention to what events the agent witnesses. Such, purely nonmentalistic accounts have been heavily criticized in the literature, pointing out that infants can make predictions even in situations they have never encountered before, thus for which they could not have acquired a rule, that there is no ground to assume that some of the proposed effects actually exist, and, most importantly, that these accounts cannot provide a parsimonious explanation for the ease with which infants (and adults) are able to predict others' behaviour in the large variety of social situations they do (Christensen & Michael, 2016; Scott & Baillargeon, 2014). Recent studies showing that the early findings replicate under conditions that rule out the postulated low-level processes, for example when the timing differences are equated in Kovács and colleagues' (2010) study (El Kaddouri et al., 2020), when behaviour rules are not applicable in the situation (Kovács et al., 2021; Meert et al., 2017) or when the agent wears transparent but not opaque googles (Furlanetto et al., 2016; though for a nonreplication see: Conway et al., 2017) or is blindfolded (Seow & Fleming, 2019) in the dot-perspective-taking task, seriously question that these low-level processes alone could account for the observed effects. Nevertheless, they do not exclude the possibility that they might play some, or in some cases even important role, for instance by directing attention to what the agent perceives (see Holland et al., 2021, for the role of directional cues in the results of the dot perspective-taking task).

1.3.2 Multi-step mentalistic accounts

A number of authors tried to reconcile the early findings, showing that children start to pass standard false belief tasks only around four years of age, with the more recent ones, indicating false belief understanding even in preverbal infants, by proposing that very young children rely on ToM capacities

that, although ensure some understanding of others' mind, are much more limited than those possessed by adults or older children. For example, focusing on the evolutionary role ToM skills may fulfil, Tomasello (2018) argues that, while four-year-old children rely on humanspecific social-cognitive skills evolved to enable cooperative interactions, to pass false belief tasks, infants succeed on the nonverbal versions of these tasks by using social-cognitive abilities that were evolved to ensure success in competitive scenarios and are already present in nonhuman primates. These include the capacity to imagine and track what the other agent sees, believes or knows, something even great apes seem to be capable of (Karg et al., 2015; Krupenye et al., 2016), but do not incorporate the ability to coordinate different mental perspectives with each other. Infants (and great apes) do not even understand that other mental perspectives, such as 'objective reality', exist. What they do, according to Tomasello, is the encoding of others' mental states without comparing those to their own knowledge of the objective situation. Young children come to understand that mental perspectives – the agent's, their own, and the objective view - may differ in the following years, via taking part in communicative and cooperative social interactions that require the coordination of mental states with others. In a similar way, Southgate (2020) argues that infants are able to track the content of other agents' (false) beliefs, partly because they lack a competing self-perspective. More specifically, the author proposes that human attention is biased towards the targets of other agent's attention from birth, and, in infants, this 'altercentric bias' is facilitated by the initial absence of self-representation. Infants can succeed on nonverbal false belief tasks, despite lacking the executive resources otherwise necessary to pass these, because, in their case, there is no need to overcome egocentric bias, i.e. inhibit their own representation of reality. They start to fail later when 'cognitive self-awareness' emerges and pass again when their inhibitory capacities become mature enough to overcome the pull of their own perspective. While both Tomasello's (2018) and Southgate's (2020) account is appealing, as they both offer an explanation not only for the question why the observed gap exists between infants' and preschooler's ToM abilities but also for how children overcome it, they leave open the question what mechanisms underlie adults' performance on 'implicit' ToM tasks.

1.2.3 The 'two-system account'

Other authors attempted to provide an account for a much wider range of data, including those coming from studies with adult participants. In specific, motivated by the question, how theory of mind can fulfil its two main roles in daily life, to be (i) efficient enough such that it can enable the fast prediction of behaviour online interactions require and to be (ii) representationally flexible enough to allow for

the attribution of a possibly indefinite number of, potentially infinitely complex mental state contents, Apperly & Butterfill (2009) proposed that, just like in number cognition, there are two systems involved in ToM, that are fundamentally different in nature, follow a different developmental trajectory and operate in parallel in adults. One of them is an early-developing, fast and efficient system that underlies performance on nonverbal ToM tasks and supports the online tracking of mental states later in adults. The other one is a late-developing, slow and effortful system, deployed when people have to reason deliberately about the mental states of others. Importantly, the authors claim that, as propositional contents are highly complex, only this latter system can represent beliefs *as* propositional attitudes. The early-developing or 'minimal' mindreading system, as Butterfill and Apperly (2013) call it later, employs a distinct set of concepts, 'belief-like states', that simply encode relational information between agents, locations and objects. Although such relational attitudes enable the generation of rough predictions of others' object-related behaviours, in a fast and efficient manner, i.e. allow the system to operate much like informationally encapsulated modular systems do, this comes at certain costs. First, the minimal mindreading system has important limitations with respect to what kind of information it can handle: it permits the tracking of beliefs about locations but not about complex combination of properties or quantifiers. Second, as belief-like states do not represent the aspect under which the object has been encoded by the agent, the minimal mindreading system does not enable understanding that different people might think about the same entity in a different way. Accordingly, it supports Level-1 but not Level-2 visual perspective-taking - a constraint the authors consider to be one of the most important 'signature limits' of the early-developing theory of mind. Finally, the use of a distinct set of concepts allows for little information flow between the two ToM systems, resulting in a more or less independent operation where the output of one does not influence that of the other (Butterfill & Apperly, 2013; Low et. al, 2016).

Initial findings indicating the presence of the proposed 'signature limits' in adults' efficient mental state reasoning and very young children's false belief understanding, seemed to support the 'two-system account'. For example, using a modified version of the dot-perspective-taking task, with viewpoint-independent (such as 0 and 8) and viewpoint-dependent numerals (such as 6 and 9) instead of dots, to investigate Level-2 perspective-taking, Surtees and colleagues (2012) found no evidence for the automatic processing of the perspective of the avatar. Neither primary school-aged children nor adults were slower in judging their own perspective when the number appeared differently for the other compared to when it appeared in the same way for the avatar as for the participants (see also: Surtees, Samson & Apperly, 2016). In line with this, some studies demonstrated a striking failure of young children (as well as adults) to take into account another agent's false belief about an object's identity, when anticipating or interpreting the other's actions, despite having no problem with attributing false beliefs to the agent about object locations. In specific, these studies found that two-to-four year old

children (and even adults) fail to understand that an agent, who saw the hiding of an object under one aspect (as a blue toy or a bunny) then the transfer of the object to another location under the other (as a red toy or a carrot), and is unaware of object's dual appearance/identity, will search for the object at its original location (Fizke et al., 2017; Low & Watts, 2013, but see: Kulke, von Duhn et al., 2018 for a failed replication of Low & Watts' findings).

Early results showing that two- and three-year-old children, who anticipate correctly in the false belief task, nevertheless respond incorrectly (but, at the same time, with high confidence) when asked to make explicit predictions about the protagonist's future actions (Clements & Perner, 1994; Ruffman et al., 2001), were taken as evidence for the second claim of the account: that the two ToM systems operate independent of each other, with no direct communication between the two. More recent results demonstrating no within-child correlations between three- and four-year-old children's performance on standard false belief tasks and certain (albeit currently highly criticized) measures of a nonverbal ToM (Grosse Wiesmann et al., 2017) seemed to provide further support for the claim.

1.4 Towards a representationally flexible and efficient theory of mind

1.4.1 The empirical criticism of the two-system account

Recent findings, showing that adults (Elekes et al., 2016; Surtees, Apperly & Samson, 2016) as well as primary school-aged children (Elekes et al., 2017) compute the content of another person's Level-2 visual perspective in interactive versions of Surtees and colleague's (2012) 'number-perspective-taking' task, pose serious challenge for the two-system account. While some authors found that altercentric interference effect emerges only if the other person focuses on the same feature of the stimuli as the participant, i.e. performs the same number-verification task (Elekes et al., 2016), others reported similar findings even if the confederate paid attention to another aspect of the stimuli (surface features instead of the magnitude of the number), that is even in the absence of a shared task goal. The effect was present as long as the other person was actively involved in the task (Surtees, Apperly & Samson, 2016). Some studies found that the way another person perceives a stimulus, e.g. upside down or close to its canonical form, may have an impact on performance even if the confederate is passive (Freundlieb et al., 2018; Ward et al., 2019), suggesting that, unlike how the two

system account predicts, Level-1 and Level-2 visual perspective-taking may operate in a similar manner in adults, i.e. involuntarily and outside awareness.[2]

With respect to young children, recent studies found that 14-month-old infants' behaviour is modulated by another agent's false belief about an object's identity, indicating that previous failures might have reflected limitations of the paradigms or measures used and not that of infants' ToM abilities. In specific, 14-month-old infants search longer if another agent, unaware of the fact that the object that has been removed from the scene (under one aspect) is actually the same as the one that was previously hidden (under the other aspect), mistakenly believes that there is still an object present (Kampis & Kovács, 2022). They also seem to be able to infer the agent's preferences based on the agent's false belief about a deceptive object's identity, anticipating in an unexpected identity-change task accordingly (Buttelmann & Kovács, 2019). In addition, a number of studies have demonstrated that infants and toddlers form much more sophisticated expectations regarding other agents' future behaviour than what 'relational attitudes' would permit, combining the agent's beliefs about a hidden object's identity (they could infer via deduction) with the agent's preferences when making predictions how the agent will act (Cesana-Arlotti et al., 2020) or inferring which object an agent intends to refer to, based on the agent's false belief about the respective object's location (Southgate et al, 2010; replicated in: Király et al., 2018; though for nonreplication see: Dörrenberg et al., 2018). Moreover, they can ascribe a belief to another agent about an object's location even if they themselves do not know where the object has been hidden, i.e. when they cannot form associations between agents, locations and objects (Kovács et al., 2021).

While the apparent absence of the proposed signature limits provides evidence against the 'two-system account', by showing that the two may be actually indistinguishable in this respect, other results directly question the main assumption of the theory, that verbal/'explicit' and nonverbal/'implicit' ToM tasks are subserved by two distinct 'theories of the mind'. In specific, a recent study found no difference between the explicit and the implicit version of Kovács and colleagues' (2010) object detection paradigm, in the extent to which the agent's belief about the presence of the ball facilitated adult participants' performance on the task, suggesting that, in the explicit version, when they had to track the agent's beliefs to be able to respond to occasional catch questions about it, adults relied on the same computations they performed when they were tracking the agent's beliefs spontaneously (Nijhof et al., 2016). In line with this, follow-up neuroimaging studies indicated the involvement of the same brain regions in the 'explicit' and in the 'implicit' (original) version of the task,

[2] It is worth noting, that in many of these tasks participants might have computed the other agent's perspective content because the presence of the confederate might have triggered thinking about his/her role in the task. Acknowledging this possibility, most of the cited authors talk about 'spontaneous' (and not automatic) visual perspective-taking, which is, nevertheless, still unconscious and unintentional.

specifically the activation of the right temporo-parietal junction and the medial prefrontal cortex (Bardi et al., 2017; Nijhof et al., 2018), brain areas, consistently found to be involved in verbal/'explicit' (Frith & Frith, 2003; Saxe, 2006), as well as in various nonverbal ToM tasks (Hyde et al., 2015; Kovács et al., 2014; Naughtin et al., 2017). Other findings, showing that the neural timing of trait inferences (as measured by the onset of event-related potentials) is independent of whether participants receive an explicit instruction to perform such inferences or not (Van Overwalle & Vandeckerhove, 2013), also suggest a shared neural basis of deliberate and unintentional mental state reasoning, in general.

Taken together, results from these two lines of research, targeting infants on the one hand and the neural bases of these abilities on the other, suggest that mental state attribution relies on the same core mechanism(s), that seems to be the same irrespective of whether the attribution takes place voluntarily or not, of the attributed content, as well as the age of those who engage the process (adults versus infants).

1.4.2 Theoretical considerations: the changing notion of automaticity

Besides the empirical findings questioning the two-system account, there is one more, fundamental issue with Apperly and Butterfill's (2009) proposal, that has to be pointed out: the core assumption that there is an inherent tension between flexibility and efficiency, i.e. that a process is either automatic but inflexible or flexible but controlled, may be in fact flawed, as it derives from a widely spread but outdated conceptualization of automaticity as an 'all-or-none' feature of cognitive processes.

Automaticity is defined in a variety of ways in the literature. Initially, processes were considered automatic to the extent they operated independent of attention, with purely automatic processes, drawing on little or no attentional resources on one end of the continuum, and nonautomatic or controlled processes, requiring substantial amount of attention, on the other. Later the two endpoints started to be conceptualized not just as two opposite but as two distinct modes of processing, characterized by a set of features (Moors & De Houwer, 2006). Automatic processing was conceptualized as unintentional, uncontrollable, efficient and unconscious, assuming a perfect correlation between these four characteristics. The core assumption was that if a process takes place more or less independent of attentional resources ('control processes') this essentially means that, on the one hand, it is activated by stimulus input, hence requires no intention from the person to engage in it, on the other hand, that it is effortless and efficient in a sense that it can run parallel with other automatic and control processes, without any interference. Therefore, it is relatively fast, compared

to controlled processes which have to be executed in a serial manner. At the same time, being independent from attentional control also means that the process runs outside awareness, and being stimulus-driven that, once the process started, it is difficult to be stopped or altered, i.e. it is uncontrollable. Accordingly, a process was viewed as either automatic, possessing all four features, or controlled, that is intentional, controllable, inefficient and conscious – something qualitatively different (Bargh, 1994; Melnikoff & Bargh, 2018; Moors & De Houwer, 2006). Importantly, some authors related the mode of processing to the structure of the cognitive architecture underlying the particular process. In particular, some considered automatic processing to be the result (and at the same time the proof) of the fact that the underlying mechanism is informationally encapsulated, modular in nature, i.e. has no access to and is not affected by information in other parts of the cognitive system (see e.g. Fodor, 1985).

The dichotomic conceptualization of automaticity has been subject to serious criticism in the last few decades, for two main reasons. First, the defining features are themselves complex dimensions, which makes it rather difficult to characterize processes along them. A process may be unintentional in some sense but not in the other, unconscious from some perspective but not from the other or efficient in some context but inefficient in another. For instance, an act or a process is considered unintentional, if it is not caused by the goal to engage in it (Bargh, 1994). This does not mean, however, that the process is necessarily stimulus-driven. Its occurrence may depend on a remote or overarching goal. For instance, one may not have the specific goal to track others' mental states but may have the general goal to track the conversation one is observing or to win a contest, which may even be conscious, as it has been pointed out by many (see e.g. Moors & De Houwer, 2006). In a similar way having a conscious goal – whether a proximal or a distal one - does not mean that the person is aware of the ongoing process. Being aware of the cause or trigger of a certain cognitive process, its effect on one's judgements, and the process itself, are three different subdimensions of consciousness that may vary independent of each other. In addition, being unaware of a certain piece of information (an input or an output) may also mean at least two things: that the information is structurally inaccessible, or it is potentially accessible, in case attention is directed at it at some point, with accessibility, again, not necessarily meaning that the information can and will be verbally reported (Melnikoff & Bargh, 2018; Moors & De Houwer, 2006). The second, more important issue with the approach is that the defining qualities of automatic and controlled processes are not mutually exclusive. In fact, in the last few decades, a vast number of studies has shown that they can co-occur in almost any combination: an act or a process can be, for example, intentional yet unconscious, like the execution of overlearnt skills, unintentional but still effortful, in a sense that it is attenuated by cognitive load, such as word reading, or unintentional yet controllable by one's goals or motivational states, like motor mimicry which can be suppressed (for reviews see: Bargh, 1994; Melnikoff & Bargh, 2018). In accordance with these

results, several studies have shown that even perceptual systems, the most paradigmatic cases of modularity, are not cognitively impermeable. Abstract, conceptual knowledge as well as the individual's goals and motivations influence even the earlies stages of visual processing, in a top-down manner such as scene segmentation or the selection of which stimuli or which features of an object get rapidly processed (Gilbert & Li, 2013), demonstrating that even processes widely considered as purely stimulus-driven are not entirely uncontrollable and inflexible.

Therefore, one may wonder whether the automatic-controlled distinction could or should be applied to mentalizing; since the two modes of processing are not mutually exclusive, there is no need to presuppose two distinct systems to explain how mental state attribution can be efficient on the one hand and flexible on the other.

1.4.3 The unity of implicit and explicit theory of mind

Empirical findings indicate that the features associated with automatic and controlled processing indeed co-exist in mentalizing. Results showing that, the presence of the altercentric interference depends on the belief whether the avatar can or cannot actually see the dots in the dot-perspective-taking task, both in adults (Furlanetto et al., 2016) and, in competitive contexts, even in nonhuman primates (Karg et al., 2015), suggest that the computation of the other agent's Level-1 visual perspective may be affected by top-down processes. In other words, Level-1 visual perspective-taking is cognitively permeable, even though the process is fast and efficient (for a detailed discussion see: Westra, 2017). Other studies indicate that the tracking of other agents' (false) beliefs may depend on executive resources, and thus may be effortful despite being unintentional. For example, Yott and Poulin-Dubois (2012) found a strong correlation between 18-month-old infants' success on an inhibitory control task and their looking time in a nonverbal ToM task (but see: Grosse-Wiesmann et al., 2017 for opposite findings with three- and four-year-old children and another paradigm). Others demonstrated that cognitive load diminishes the effect of the other agent's false belief on adults' looking behaviour in anticipatory looking tasks (Schneider, Lam, et al., 2012). Furthermore, if a secondary task taxes working memory, it impairs even the 'automatic' processing of an avatar's Level-1 visual perspective in the dot-perspective task, as indicated by a marked decrease in altercentric interference effect in the dual-task compared to the no secondary task condition (Qureshi & Monk, 2018). Finally, findings discussed in the previous section, indicating a sophisticated understanding of false beliefs in infants and rapid computation of not just others' Level-1 but also their Level-2 visual

perspective in an unintentional, but seemingly context-sensitive manner, in adults, indicate that fast and efficient mental state attribution is far from being inflexible.

Although some kind of a complementary trade-off may exist between the efficiency and flexibility (for instance in terms of how complex a rapidly computed content can be), such results clearly show that, contrary to previous assumptions, the two are not mutually exclusive features of mental state attribution. People seem to be able to ascribe mental states in a manner that allows for smooth interactions, with no strict limitations on the type of contents that can be attributed this way. Therefore, one might argue that is no need to postulate two distinct ToM systems to explain how these two requirements of ToM can be fulfilled in the daily life of humans.

In line with these considerations, several authors have suggested the existence of a single 'mindreading system,' behind all forms of mentalizing (Carruthers, 2017; Kovács, 2016; Leslie et al., 2004), proposing different solutions for how efficiency and flexibility can be achieved simultaneously within such a framework. Carruthers (2017), for example, argued that there is one ToM system, with one set of concepts and inferential rules, that, however, can operate in various ways: sometimes automatically, sometimes in conjunction with a goal, sometimes closely together with domain-specific or domain-general executive resources, according to the needs of the specific situation. Leslie and colleagues (2004) put forward the existence of a learning mechanism, made up of two components that work closely together during the process of belief-desire reasoning: the 'Theory of Mind Module' (ToMM), the core representational system enabling the attribution of beliefs and desires, that is triggered automatically by the presence of another agent and is responsible for quickly identifying plausible candidates for the content of the other agent's belief, and the 'Selection Processor' (SP), domain-general control processes that help to select the correct response from those 'offered' by the ToMM, via inhibiting the default response when necessary. In this view, efficiency is implemented by the way ToMM operates while flexibility is ensured by the SP. Kovács (2016) took another approach, proposing a new type of representational structure, the 'belief file', to solve the problem, which would allow for efficient encoding and flexible handling of others' belief content via the format it takes. In specific, it would make such a functioning possible by virtue of having a structure with two variables (placeholders), one for the agent and one for the belief content, that can be accessed and manipulated independent of each other, enabling the fast tracking of changes. Finally, even others postulated the existence of a multi-system architecture, made up of distinct elements, which nevertheless operate in a tightly integrated manner to construct situational models for the interpretation and prediction of the behaviour of others (Christensen & Michael, 2016).

At the same time, most of the authors abandoned the notion of automaticity and switched to the term *spontaneous* instead (Elekes et al., 2016; Freundlieb et al., 2018; Nijhof et al., 2016; Surtees, Apperly & Samson, 2016), to describe how mental state attribution takes place in online social interactions,

thereby acknowledging that the mere presence of an agent may not be a sufficient trigger for the process on the one hand and that it may not be completely independent of attentional resources on the other. Throughout the thesis we will also use this term, to refer to processes that take place unintentionally, taking no strong position on whether the particular process possesses one or more of the features associated with automaticity, i.e. whether it is effortless, uncontrollable and completely unconscious.

1.5 Setting the problem

To summarize, theoretical considerations, as well as empirical findings suggest that flexibility and efficiency can co-exist simultaneously in the mentalizing of human adults (and, possibly, even in infants' ToM), thus there may be no need to presuppose two distinct ToM systems to fulfil these two roles, as suggested by the two-system account. If humans indeed possess a single, efficient yet representationally flexible theory of mind mechanism, then one may argue that they should be able to spontaneously perform all those ToM computations that successful navigation in the social world requires, independent of the complexity of the to-be-computed content or the processes involved. Crucially, humans participate (and, thus have to be successful to a considerable degree) in a much wider range of social situations than those that have been extensively investigated in the past (to establish whether or not performance on the implicit ToM tasks reflects the genuine understanding of others' mind). Whether the capacity to represent others' mental states emerged to aid competition for resources, specifically the manipulation of other group members, as some authors claim (Byrne, 1995) or to support the various forms of collaboration co-existence in groups requires, as many others proposed (see e.g. Tomasello et al., 2005), the evolutionary function of theory of mind is to help the prediction and interpretation of other agents' behaviour, and thereby enable flexible adaptation, not only in false belief type of scenarios but in all kinds of social situations humans face in their everyday life.

Importantly, most of the situations humans encounter in their daily life differ from the scenarios covered by the various versions of the standard false belief task, in many respects, including (1) their dynamics, i.e. the number and type of the events witnessed by the observer (and, consequently, whether or not they require the revision of beliefs and expectations); (2) the kind of beliefs agents hold, e.g. whether they have a firm belief or merely hypotheses about the actual state of the world; as well as (3) the sources of those beliefs, e.g. whether they are based on what the agent witnessed or

on what the other could merely infer, via deduction. Hence, they differ both in the type of contents observers have to compute and in the computations they have to perform to be able to act appropriately or prepare for the other agents' (potential) future actions.

Regarding the dynamics of social situations, the environment often changes, and it does so repeatedly and rapidly, in a way that it affects the content of other agents' mental states, which often forces people to revise their assumptions regarding what others see, believe or know, fast enough to be able to swiftly adapt to the newer and newer situations. In a similar way, other agents' behaviour may also change rapidly, sometimes rather unexpectedly. Importantly, in this case the person first has to come up with an explanation for the change in the other's course of actions before re-attribution and the subsequent, appropriate behaviour adjustment could occur: realize that the previously attributed content has become outdated or his/her original assumptions might have been wrong, then 'reverse engineer' the other's mind on the basis of the observed behaviour. For instance, upon seeing someone stepping down the road while a truck is loudly approaching, one must realize that the other did not notice the danger, probably due to listening to music, and act accordingly, grabbing the other's clothes instead of shouting.

With respect to the possible contents one can attribute, agents who lack a certain piece of information, for example, due to not witnessing certain events, such where their favourite toy was hidden, rarely hold a false belief about the situation. They are rather uncertain about the current state of affairs and entertain multiple hypotheses about what might be the case ('it may be in the bed, in the toybox or under the sofa'), assigning a certain level of probability to each of those. This results in a belief the content of which is more complex than the ones observers have to attribute to the other in false belief scenarios, i.e. is composed of several elements, possibly with a disjunctive relationship between those. Such belief contents have two other features that are worth consideration. First, if the other's hypothesis space also includes the actual location of the toy, one can argue that this content is neither completely true, nor completely false. Second, the representation of such a content, i.e. the other's 'hypotheses' does not enable the prediction of the other agent's behaviour with the level of certainty simple true or false beliefs do, in the sense that if, for instance, the alternatives the other upholds are equally probable, it is difficult to predict where s/he will search first for the toy. To emphasize this latter aspect, we will refer to these belief contents with the term 'underspecified' in the followings. Despite this limitation, i.e. the fact that they do not make the other agent's behaviour fully predictable, the representation of such belief contents may nevertheless be useful, as it makes it possible to restrict the range of actions one can expect from the other, and in this way, to 'prepare' for the future.

Finally, regarding the sources of beliefs, so far ToM studies have almost exclusively investigated situations in which the other agent's belief was based on what he/she has witnessed or has been informed about via communication. Crucially, however, agents might also have beliefs based on what

they have inferred deductively from what they have witnessed or were informed about and act on the basis of these conclusions. For example, seeing an open entrance door upon coming home and the shadow of a man on the wall, a person, who did not notice her husband's bag on the floor, may infer that a burglar broke into the house. Tracking what information (or premises) another person has, and representing the conclusions they may draw from those, may allow the observer to interpret why the other person starts screaming and predict the other's next move, that he/she will run out of the door. Thus, just like the representation of the other's hypotheses space in other situations, it enables the prediction of (and the preparation for) a much wider range of actions than the mere representation of what the other knows would allow for.

Despite the fact that these situations are relatively common, little is known about how humans perform the mental state computations successful adaptation in such cases requires. There is some evidence that infants update the content of the attributed belief and revise their expectations, upon witnessing that the agent has received a possibly relevant piece of information (Song et al., 2008; Tauzin & Gergely, 2019) and toddlers do the same if they themselves learn that their previous assumptions have been wrong (Király et al., 2018). These suggest that these processes take place spontaneously, but little is known about how such an update takes place if the person has to realize that there is a need for belief revision from the other agent's observed actions. Some findings suggest that infants and toddlers may also track the inference another agent may draw from a misleading piece of information (Song & Baillargeon, 2008). They may also understand that an agent who has not witnessed a hiding is uncertain about the object's location, expecting the agent to search randomly at the two locations (Knudsen & Liszkowski, 2012b). However, the interpretation of these findings, i.e. what is represented by the infants, is far from clear. There are no adult studies either that would target these issues, leaving open the question what the nature of these processes is.

Given that the aforementioned computations can be considered more complex than the attribution of simple true or false beliefs about the identity or the location of objects, as they involve multiple steps and/or the attribution of a content made up of more than one element, one may argue that they only take place when it is absolutely necessary, voluntarily and with much effort. Nevertheless, if ToM relies on a mechanism (or set of mechanisms, functioning in an integrated manner) that is indeed both flexible and simultaneously efficient, evolved to support success in a variety of online social interactions, human adults should be able to perform them in a way that enables swift adaptation to the other agent's behaviour, at the given moment or when this becomes necessary in the future. At minimum, such attributions should take place *spontaneously*. Crucially, this does not mean that any of these computations should be completely effortless, fully unconscious or stimulus-driven, independent of the influence of the variety of factors that may affect whether people engage in mentalizing in a certain situation, which may range from the perceived relevance of the agent to the

availability of attentional resources. All this means is that they should take place unintentionally, without the need for any specific goal to perform the particular computation, i.e. independent of any overt task or external prompt. The present thesis aims to investigate this assumption. In particular, it aims to explore whether the three ToM computations we identified as playing an important (possibly key) role in how well human adults can adapt to others in a number of everyday life situations - namely 1. updating other agents' mental states, based on the behaviour they demonstrate in a situation, 2. encoding the hypotheses they likely entertain and 3. representing the conclusions they may draw from the beliefs they hold – take place spontaneously, even when this would not be necessary in the given situation, thereby enabling smooth adjustment and/or fast reactions to others' behaviour at later timepoints when adaptation becomes inevitable or useful. In the following three chapters we present empirical work that addresses these issues.

Chapter 2 presents two eyetracking experiments in which we investigated whether human adults spontaneously update the content of another agent's mental state and revise their expectations regarding her future behaviour upon observing the other repeatedly acting in a way that is incongruent with their initial assumptions regarding what she sees or knows. To test this question and to gain insight into how the process unfolds we analysed how participants' anticipatory looking behaviour and reaction times change over the course of trials following the other's first unexpected action. In addition, we also investigated whether human adults generalize what they have learnt about the other agent and use this knowledge to predict her actions in subsequent interactions. *Chapter 3* reports results from five experiments that used a change detection paradigm to explore whether people represent the content of another agent's hypothesis space spontaneously, even if this is not necessary for the task they perform. More specifically, we tested whether in a situation where another agent is uncertain about where an object has been hidden, hence represents two, equally likely alternatives regarding its location, this 'underspecified belief content' of the other affects the way participants allocate their spatial attention, and via this, their sensitivity for changes at different locations. *Chapter 4* presents four online experiments designed to investigate whether human adults represent what conclusions another agent may draw from the beliefs she holds (regarding an object's identity or location), spontaneously, or whether they do it only when they have to track this information. In particular, we tested whether a potential conclusion another agent may arrive at (that is different from the conclusion the participant can draw) modulates adults' estimations about the probability of certain outcomes (where or what an object is), and the time necessary to perform those estimations, in situations where the other ends up considering more outcomes 'possible' than participants do, as a result of lacking a certain piece of information. Besides our main research question, we also investigated the scopes and limitations of this capacity, by using different scenarios in the different

experiments we ran, which required participants to perform logical inferences of varying complexity (both from first- and third-perspective).

In all three studies, we investigated situations in which the other's belief was irrelevant for the task participants had to perform and the other agent was either merely present or acted but the participants' success did not require the prediction of his actions. Some experiments yielded positive, others mixed or rather negative results under such circumstances, providing important insights into what may be the minimum necessary preconditions for the three different ToM computations to take place in human adults. Since in all three studies we used novel paradigms and measures to address the research questions we had, the findings also provide useful information on what methodological approaches may be fruitful in their exploration in the future. The last chapter summarizes these 'insights' and analyses what our results tell about the broad research question: how the functioning of mature ToM contributes to human adults' flexible adaptation to the social world.

Chapter 2: Updating other agents' mental states on the basis of their behaviour

2.1 Theoretical background

From crossing the street to playing basketball or taking part in an everyday conversation, efficient social interactions require making correct predictions about what others will do or say. It has been widely accepted that humans perform such predictions by taking into account the unobservable mental states of other agents, expecting them to act in line with what they want, see, believe or know. If, for example, in a basketball game, the lead player believes that the teammate closest to her did not see the opponent on the right, she will take this into consideration when formulating expectations regarding the teammate's future move and will adjust her own movements accordingly. In a similar vein, if players observe their team leader deviating from the strategy that the team has agreed upon, they will assume that the leader has a good reason to do so, even if the exact reason for her behaviour might not be that straightforward at that very moment they notice the deviation. Furthermore, the explanations they will come up with to interpret the situation will likely involve reasoning about the mental states of the leader, endorsing a discrepancy between her current mental state and their own, as well as from the mental state they attributed to her before (e.g. "maybe she misunderstood what we ought to do'). How humans make such rapid and seemingly complex inferences, i.e. revise their beliefs about the knowledge state of others, and, consequently, their expectations regarding others' actions upon encountering an unexpected behaviour has been so far largely unexplored.

To date, most of the research investigating ToM abilities has focused on the question how children and adults make inferences about others' mental states and take those into account when predicting their behaviour. Studies from the last 15 years indicate that people compute the content of other agents' mental states spontaneously. As discussed in detail in the previous chapter, their responses are affected by what another agent sees and knows, as well as by how an object is seen from an interaction partner's point of view in tasks that do not require the tracking of others' mental states (Buttelmann & Buttelmann, 2017; Elekes et al., 2016; Kovács et al., 2010; Samson et al., 2010; Surtees, Apperly & Samson, 2016; van der Wel et al., 2014). They also display eye-movement patterns indicative of attributing a false belief to another person in nonverbal versions of the false belief task, even when they are not instructed to predict the other agent's actions. In specific, they look more towards the location where the agent falsely believes the object to be hidden than at the location corresponding to the true state of affairs, right before the agent would start acting, without being aware of doing so and even after repeated presentation of the same stimuli, suggesting that they spontaneously anticipate the other to search based on her belief (Schneider, Bayliss, et al., 2012; Schneider et al., 2014). Recent failures to replicate the original findings (Burnside et al., 2018; Kulke, von Duhn et al.,

2018) and to demonstrate any sign of spontaneous action anticipation in such tasks (see e.g. Schuwerk et al., 2021), led many authors to question this interpretation, as well as the validity of anticipatory looking as a measure of false belief understanding, in general. However, other results suggest that these failures may reflect limitations of the specific paradigms, for instance, the limited efficiency of the triggers used in these studies in eliciting spontaneous action prediction in adults. In particular, eyetracking studies using different versions of the visual world paradigm, in which people hear utterances referring to some element of the visual display they watch, consistently report an early sensitivity of adults to others' false beliefs and discrepant visual perspectives, albeit in the face of a strong egocentric bias. Specifically, in situations where the cognitive load is not too high, a looking behaviour of listeners indicates a spontaneous consideration of the other's mental state while they are processing utterances describing where the other will look for a hidden object in nonverbal false belief tasks (Symeonidou et al., 2020) or while listening to instructions regarding which object to move in a display in referential communication tasks (Cane et al., 2017; Hanna et al., 2003).

Importantly, the smooth unfolding of social interactions requires not only the representation of other agents' mental states, but also a fast and flexible updating of the represented content, whenever new evidence suggests that one's original assumptions regarding what the other believes or knows might no longer hold – they have become outdated or might have been wrong from the beginning of the interaction. Such situations arise, for example, when one observes another agent witnessing an event that should provoke a change in her knowledge state (e.g. the pedestrian turning her head right and noticing the approaching truck), or when one acquires a new piece of information that is incompatible with the earlier attributed mental state (e.g. one thought that the pedestrian saw the truck but now she behaves as if she did not, stepping down the road to cross it). Notably, these two situations differ markedly in how updating is performed. In the first case, one has to perform the update based on the new information that became available to the other, in a prospective manner, and conclude that the other's mental state has changed as a result of the observed events. The second situation, on the other hand, requires one to reconsider one's own assumptions and recompute the initially attributed content retrospectively, on the basis of indirect evidence suggesting that one might have made an incorrect attribution previously.

Developmental studies indicate that the ability to update another agent's belief in a prospective manner, specifically based on communicated information, is already present by 13 months of age (Song et al., 2008; Tauzin & Gergely, 2019), suggesting that such computations might take place spontaneously in humans. In specific, infants look longer if an agent looks for the hidden object at the wrong location after observing a communicative interaction with a knowledgeable other, indicating that they expect the knowledgeable other to communicate relevant information (that corrects the agent's false belief) and the agent to take that into account when performing the search. Much less is

known about how people update other agents' mental states when the situation requires retrospective recomputation of the previously attributed belief content. A recent study suggests that, by their third year, children can correct their earlier attributions (whether or not an agent has a true belief about the location of an object), spontaneously, when provided with new contextual information that sheds new light on what events the agent could have witnessed before (for example when it turns out that the sunglasses that an agent wore while a location change happened were not transparent, hence she is not aware of the new location of the objects; Király et al., 2018). Such findings corroborate results from studies investigating the process of first-person belief revision in adults, which show that humans prioritize observed data over previous assumptions in conditional reasoning tasks when they encounter a new piece of information that contradicts their prior conclusion, rejecting the major premise (i.e. the regularities expressed in the conditional) rather than the antecedent or the new evidence in such cases, to resolve the inconsistency (Elio & Pelletier, 1997).

Crucially, situations that necessitate the retrospective update of another agent's mental state arise not only when one receives a new piece of information from others indicating that he/she might have been wrong regarding what the other agent knows. In fact, it is much more common that one has to realize the need for such a revision from the way the other agent behaves in a given situation. Failures to correctly predict the other agent's actions constitute one of the most important signals that one might be mistaken about the other's goals or knowledge states, especially if they arise in an environment that is otherwise stable and predictable. Although such prediction errors are quite common in everyday life, to date the nature of the mechanism that allows the interpretation of and adaptation to other agents' unexpected actions, has been largely unexplored.

Notably, to retrospectively update the mental state previously attributed to the other agent on the basis of his/her unexpected action(s), one first has to realize that there is a need for such a revision, then has to find an adequate explanation for the observed behaviour, i.e. perform two additional steps compared to other forms of mental state update, before the recomputation itself could occur. Since events that violate expectations induce ambiguity and it is well known that ambiguity prompts organisms to engage in activities that minimize it (Courville et al., 2006; O'Reilly, 2013), one might argue that the unexpected actions of the other agent may spontaneously trigger reasoning about the underlying causes. Such a reasoning likely entails the generation of multiple candidates, before the inferential process would end up with selecting the one that seems to be the most likely (or at least a 'good enough') reason for the observed behaviour in the light of the available information. If the observer's higher-order expectations about the plausibility of errors in the given context make it unlikely that the observed behaviour reflects a simple failure to properly execute the intended action, it can safely be assumed that the reasoning process ends up with updating the content of the mental state attributed to the other agent before. Considering humans' remarkable capacity to infer the goals,

beliefs and stable preferences of any agentive entity, merely based on the actions (i.e. simple movement patters) it performs and the environmental constraints of those actions (Baker et al., 2009; Baker et al., 2017), they should also be able to efficiently update previously attributed beliefs on the basis of such information.

Everyday intuition suggests that people update other agents' mental states in a fast and efficient manner upon observing an action that violates their expectations - without engaging in conscious, slow and effortful reasoning prior to the adjustment of their own behaviour. They quickly intervene if someone behaves as if being unaware of an imminent danger (e.g. steps down the road despite an approaching car), provide additional information if an interlocutor has difficulties in interpreting a question, and swiftly come up with possible interpretations why a person could have deviated from the expected behaviour. It is not entirely clear though whether adults perform such computations spontaneously, whenever they encounter an unexpected action, or only when it is necessary, to avoid harm or the breakdown of an ongoing interaction.

A previous study indicates that people might indeed update other agents' mental states spontaneously, upon observing a behaviour that contradicts their assumptions regarding what the other knows, even if their task does not require prediction of the other agent's actions. In particular, using a referential communication task, Rubio-Fernández (2017) found that most of the participants, who played the role of the follower, noticed when their partner's verbal description of the target did not match her alleged knowledge state (i.e. what she could see according to the information they received) and inferred that sometimes she might have seen more of the visual display than they were told. Importantly, by the end of the task, these participants tended to look more towards the object that was marked as not visible to the other, suggesting that these inferences were drawn spontaneously, upon noticing the unusual behaviour. Although these findings indicate spontaneous updating of the other's perspective, since the study was not designed to investigate the retrospective updating of other agents' mental states, the findings do not allow firm conclusions to be drawn regarding how such computations are performed. More importantly, they do not shed light on how the process evolves and whether people also revise their expectations regarding the other's behaviour and adjust their own behaviour spontaneously to the other's actions.

The present study aimed at filling this gap. More specifically, it had two main goals. First, we intended to address the question whether people update the content of another agent's mental state and revise their expectations about her future behaviour, spontaneously, even when this is not necessary, i.e. they could solve their task without doing so, upon observing the other repeatedly acting in a way that is inconsistent with their initial assumptions regarding what she sees or knows. Second, we intended to explore how such a process unfolds.

To this end, we developed a virtual referential communication task in which participants were asked to react to another person, 'the partner', who performed a categorization task in another room, while their eye movements were recorded. Unbeknownst to the participants the 'partner' was a computer programme. In Experiment 1 the 'partner' first had to select one out of two pictures, depicting animals, following the auditory instructions of an unseen director, then had to categorize the picture based on the colour of its frame (blue or green), by clicking on one of two boxes, which resulted in the lighting up of the selected box. Participants' task was to simply click on the box that lit up, as fast as possible. Thus, they were not instructed to track the 'partner's' mental states or anticipate her behaviour. Nevertheless, they might have done so, spontaneously. Importantly, two of the four colours used for the picture frames were ambiguous i.e. harder to categorize. Our crucial experimental manipulation was that after a period of correct categorization, the 'partner' started to systematically miscategorize one of these two ambiguous colours, as if she had changed her mind regarding which category that specific colour belongs to. To measure whether and how participants update the other's mental state (how the partner represents a particular colour) upon observing the miscategorizations, and revise their future expectations regarding the other's actions, we recorded their anticipatory looks towards the boxes in the period preceding the other's decisions and analysed how looking behaviour changes following the first change in the other's behaviour. We hypothesized that, if participants spontaneously update the other's mental state, after the observation of some unexpected actions ('miscategorizations'), they will start to look more towards the box that is incorrect from their first-person perspective, but corresponds to the other's belief regarding the frame's colour, as compared to how much they look towards the incorrect box in the condition when the frame has the other ambiguous colour that is categorized properly by the 'partner'. Besides anticipatory looking, we also analysed changes in participants' reaction times, to gain insight into how people adjust their behaviour after revising their expectations regarding the other's actions. In specific, we predicted that, after an initial increase, reflecting surprise, participants will quickly adapt to the change: they will become significantly faster, to the extent that the difference between the speed of reactions to expected and (initially) unexpected actions will eventually disappear.

Importantly, the hypothesized changes in our measures may also reflect the revision of a rule participants had acquired before (a specific colour is associated with a specific box) and not just the updating of the 'partner's' mental state. To be able to decide which of these two accounts hold[3], we also administered an explicit perspective-taking task at the end of the experiment. In this task, participants had to categorize the pictures used in the anticipatory looking task, as well as geometric

[3] Note that there is no straightforward way to disentangle these two explanations merely by exploring how the process unfolds or how fast signs of update emerges.

shapes, having the same colour as the picture frames (used in a further task they performed – see below), either from their own or from the partner's perspective. We hypothesized that if participants can take into account how the other perceives the miscategorized colour (i.e. that their 'partner's' view differs from their own), when categorizing items having that specific colour from the partner's perspective, it can be safely assumed that prior changes in their looking behaviour (and reaction times) reflect the updating of the other's mental state and not that of a nonmentalistic rule. In addition to this task, following the experiment, we also asked participants whether they noticed anything peculiar in their partner's behaviour and, if yes, how they interpreted the behaviour of the other, to gain insight into whether and to what extent the updated content is consciously accessible.

Besides investigating whether people spontaneously update the other's mental state when witnessing a behaviour that warrants such a revision, we also aimed to address a further, related question that, to our knowledge, was not explored before. Specifically, we asked whether participants would also use this newly acquired information about the other agent's mental state spontaneously in a subsequent interaction. Given that the content of the mental state participants had to attribute in the anticipatory looking task was not an episodic information (i.e. an object's location at a specific timepoint), as it usually is in standard ToM tasks, but rather a more stable or trait-like one (i.e. how the other represents a colour), one may expect people to generalize what they have learnt about the other to future social interactions. To this end, following the anticipatory looking task, we administered an additional virtual coordination task to participants in which they could use their newly acquired knowledge about their partner (how the other sees one of the ambiguous colours), to predict where on the screen the other expects them to pass an object. In this 'implicit transfer task' participants could optimize their joint performance with the partner (to be as fast as possible to move an object to a goal location) by taking into account what they have learned about other's beliefs regarding one of the ambiguous colours in the anticipatory task. Spontaneously applying the result of the updating process outside of the specific task context where learning has occurred, to make predictions about the other's behaviour, could provide further evidence for the flexibility of the ToM system which enables humans to efficiently take part in social interactions.

Experiment 2 differed from Experiment 1 only in one aspect: there was no explicit categorization rule. Instead, the 'partner' had to make similarity judgements, to highlight the subjective nature of her decisions and thereby facilitate updating of her perspective.

2.2. Experiment 1

2.2.1 Methods

2.2.1.1 Participants

Thirty-seven university students were recruited for the experiment via a student job agency and the university's research participation system (SONA systems). Target sample size was determined based on the only known anticipatory looking study that used a multiple-trial design to investigate implicit ToM in adults by Schneider, Bayliss and colleagues (2012). The data of 7 participants was excluded either because of technical error (n=4) or because participants did not meet the inclusion criteria for the eyetracking analysis (had >30% segments with >50% missing datapoints in the anticipatory period: n=2 or did not develop correct anticipations in the familiarization phase of the anticipatory looking task[4]: n=1). Thus, the final sample consisted of 30 participants (age: M_{age}= 23.77, SD_{age}=4.34, 16 males). All of them were right-handed and had normal or corrected-to-normal vision. The study was approved by the EPKEB, Hungarian Ethical Review Committee for Research in Psychology; participants signed informed consent prior to the experiment and received monetary compensation or gift vouchers for their participation (equivalent to approximately 7 Euros).

2.2.1.2. Anticipatory looking task

Stimuli

Visual stimuli consisted of two white squares – the 'boxes'- of the same size (with the Hungarian words for 'BLUE' and 'GREEN' displayed in their middle), and line drawings of four animals (goat, pig, snail, goose), each of which subtended 4.27° of visual angle horizontally and vertically. The animal pictures were presented on a white background (the same size as the boxes), with a coloured frame (see **Figure 2.1a** for examples), subtending 0.99° of visual angle in width. Stimuli were displayed on a plain grey background.

[4] Correct anticipation was defined as a mean proportion of looking > 0.50 in the ambiguous trials (averaging the later miscategorized and properly categorized ones), either in the first or the second anticipatory period (for details see the Supplementary Materials).

The colours of the frames consisted of two shades of blue and two shades of green, which differed only in hue but not in saturation or brightness. RGB values for the colours were the following: blue: 2-147-210 and 'blueish': 2-199-208; green: 2-209-95 and 'greenish': 2-209-163. Importantly, one shade of each colour was harder to categorize (as blue or green) than the other by virtue of being closer to the other colour in the RGB colour space. The aim of this manipulation was to make it credible for participants that another person may perceive these colours in a different way. We refer to these harder-to-categorize colours as ambiguous colours and the other two colours as unambiguous colours. The auditory stimuli comprised four sentences (instructions to the partner: e.g. - Put [the target animal] into the box", in Hungarian) that were identical except the last word, which identified one of the four animals (target word; goat, pig, snail, goose, in Hungarian). All four target words were two-syllabi long, started with a consonant and fell into roughly the same frequency interval, based on the frequency information provided by Hungarian Webcorpus (www.szotar.mokk.bme.hu/szoszablya). The sound files were edited so that each target word had a duration of 680 ms. The sentences were prerecorded in a sound-proof room by a native Hungarian speaker.

Apparatus

Eye movements were recorded by a Tobii X60 eyetracker (Tobii Technology, Sweden). Gaze data was recorded at 60 Hz, with a spatial resolution of 0.2° and an accuracy of 0.5°, in a dimly lit room. To ensure good data quality we used a chinrest with a forehead support (SR Research Head Support), the height of which was set in a way to ensure a 52 cm viewing distance from the screen. Stimuli were presented on a 17-inch screen, with a screen resolution of 1280 x 1024, using Psyscope B77 software (http://psy.cns.sissa.it/). The sentences were displayed via loudspeakers, placed on the two sides of the screen. Responses were recorded using an Apple Wired Mouse, the initial position of which was fixed on the table.

Procedure

Upon arrival, participants were told that they will perform the experiment in pairs but in separate rooms, connected via Internet. They were told that, since their partner is late, unlike as usual, they will not receive the instructions together. To ensure that participants believed that they perform the task together with another human, the instructions were phrased in plural and, right after the experimenter finished the explanation of the task, participants 'heard' the arrival of the 'partner' (unbeknownst to

them a pre-recorded sound-file with a confederate). At this point the experimenter left for several minutes, allegedly to explain the task to the other. Following the experimenter's return, participants were presented with a five-point calibration sequence. If necessary, this was repeated, until at least four points were marked as correctly calibrated. Then, the instruction was repeated and the task started.

The task consisted of a familiarization and a test phase. Trials had the same structure in both phases (see **Figure 2.1b)**. Each trial started with the presentation of a central fixation cross on a white rectangular background (subtending 4.40° of visual angle in height and width) for 1500 ms. This was followed by the presentation of the two boxes, one on the left and one on the right and the verbal instruction for the 'partner' (*'Put [the target animal] in the box'*; 1282 ms)[5]. When the last word, referring to the target animal was displayed, two pictures depicting two animals (the target and a distractor) appeared on the upper and lower parts of the screen, at equal distance from the centre, with their appearance time-locked to the onset of the target word. One of the pictures always had a blue or a blueish frame, the other a green or a greenish one. The appearance of the pictures marked the beginning of the period in which the 'partner' had to select the target. This period lasted 2000 ms and was followed by a variable jitter (0 to 750 ms in 250 ms bins), to make it more credible that the variable delay after the target word reflects the decision-making process of another human. Following the jitter, a red circle appeared around the target picture, for 300 ms, indicating the partner's choice. After this the target picture disappeared and the screen remained still for 1500 ms until the 'partner' selected a box. Selection was indicated by the lighting up of the respective box for 300 ms (target box). Participants had to provide their response at this point: click on the box that lit up as fast as possible. The instruction emphasized the importance of speed, specifically that they did not have to wait until the 'flashing' of the box ends. Response period lasted for a maximum of 2800 ms or until response was given and was followed by a 2000 ms intertrial interval during which the screen was blank.

The familiarization phase consisted of 16 trials. There were four types of trials defined by the colour of the target picture's frame – unambiguous blue, unambiguous green, ambiguous blue and ambiguous green – each of which was presented 4 times, in the same pseudorandomized order for each participant (details of the counterbalancing and the randomization are presented in the Supplementary Materials). Importantly, during this phase, the 'partner' always acted as expected, i.e. selected the target picture and the appropriate box. As a result, during these trials participants could develop clear expectations regarding how the events unfold and, consequently, anticipatory looks before certain events took place.

[5] In Hungarian, the target word is regulalrly placed at the end of the sentence e.g.'Tedd a dobozba a *csigát!*' ('Put into the box the *snail*!').

The familiarization phase was followed by a short break, in which the experimenter left the room, to 'set things' for the partner. Following her return, participants were again presented with the instruction. Then the experimenter left and a 128-trial test phase started.

Test trials were identical to familiarization trials, with one crucial difference: on 25% of the trials, after the selection of the target picture, the 'partner' did not act as expected. In specific, 'she' started to systematically miscategorize one of the ambiguous colours, while still categorizing the other three colours properly. Such a manipulation yielded the following four experimental conditions: 'properly categorized blue', 'properly categorized green', 'properly categorized ambiguous' and 'miscategorized ambiguous'. The identity of the miscategorized colour (blueish or greenish) was kept constant throughout the test phase and was counterbalanced across participants.

Participants received four blocks of 32 test trials, with each block containing eight trials from each condition, resulting in a total 32 trials per condition. The distractor's frame colour, the identity of the target picture (i.e. which animal was presented) and its position (up or down) was counterbalanced within each condition, the identity of the distractor within participants, across conditions. The side of the boxes (BLUE left-GREEN right or GREEN left-BLUE right) was fixed for each participant throughout the whole task and was counterbalanced across participants.

The order of the trials within the blocks was pseudorandomized, such that there were no more than four consecutive trials with the same target picture location (up versus down), action type (properly versus miscategorized, restriction only applied in the test phase) and target box location (left versus right), and no more than three consecutive trials with the same target item (animal) and target colour frame. The blocks were separated by short self-paced breaks, during which participants were allowed to move their heads.

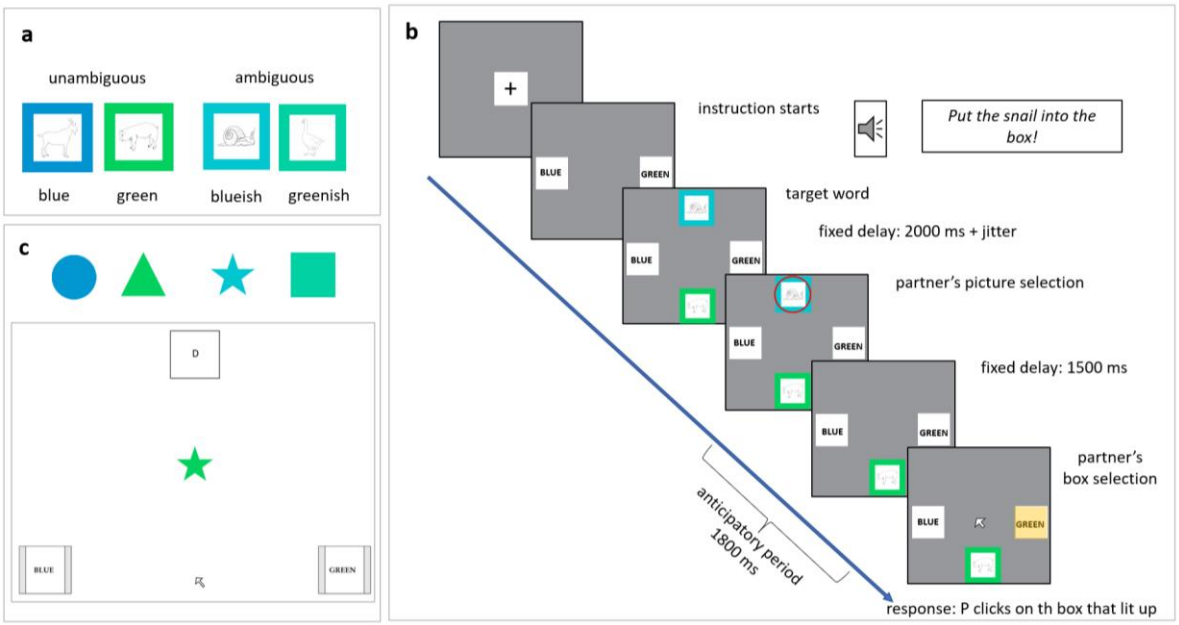


Figure 2.1. (a) Stimuli and (b) trial structure of the anticipatory looking task used in Experiment 1 and Experiment 2. The partner first selected the animal named in the instruction then categorized the picture according to the frame's colour, by clicking on one of the two boxes, which resulted in the box lighting up. The participant's task was to click on the box that lit up. The example presents a trial from the miscategorized condition. In Experiment 2, instead of the colour labels, blue and green coloured squares were presented inside the boxes. (c) Stimuli and the layout of the implicit transfer task. The participant's task was to first click on the geometric shape then on the gate where the partner waits for his/her pass.

2.2.1.3. Implicit transfer task

Stimuli and apparatus

Stimuli consisted of two white squares of the same size (the 'gates'), with grey bars at their lateral sides, and the Hungarian words for 'BLUE' and 'GREEN' in them, that were displayed at the bottom of the screen, on the right and the left, another bigger white square (the 'box') with the letter D inside (denoting the Hungarian word for 'box'- 'Doboz') presented at the top of the screen, in the middle, and four coloured geometric shapes (labelled as 'figures'). The 'gates' and the 'box' subtended 4.27° of visual angle in height and width, the geometric figures 3.30° of visual angle in diameter each.

The colours were identical to the ones used in the anticipatory looking task (blue, green, blueish and greenish). Stimuli were presented on a white background, using the same screen, resolution and software for presentation that was used in the previous task. Participants' responses were recorded with the same Apple Wired Mouse as before.

Procedure

The task started with the instructions and an example presenting the structure of the trials. Participants were told that they would participate in a joint reaction time task in which the goal was to move figures to a 'box' together with their partner, as fast as possible. Their task was to pass the item to the partner as quickly as they could, by first clicking on the 'figure' and then on the 'gate' where their partner 'waits' for the figure. The 'partner's' task was to finish the round by moving the figure into the 'box' as fast as she can. Participants were told that the two of them will receive points for each fast enough round and they were made aware of the fact that if they do not pass the figure through the gate where the partner 'waits' for it, it is unlikely that they can earn points in that given round.

Each trial started with the presentation of a central fixation cross (subtending 4.40° of visual angle in height and width) for 1000 ms, which was followed by the appearance of the two 'gates' in the left and the right bottom corners of the screen, at equal distance from the centre, and the presentation of the 'box' at the top (for the exact layout see **Figure 2.1c**). 1000 ms later a figure appeared in the middle of the screen and stayed there until the participant clicked on it or for a maximum of 3000 ms. The presentation of the figure was time-locked to the appearance of the cursor, the position of which was fixed in the lower part of the screen (vertically aligned with the 'box', at equal distance from the 'gates'). After the participant's response, the figure disappeared and participants had a maximum of 2000 ms to select one of the gates, i.e. to pass the figure to their 'partner'. If no choice was made within this time window, the trial ended. Upon clicking on the 'gate' the figure appeared inside and 'the partner's' turn started, which lasted for 2000 ms. After this, the trial ended and, following a 1000 ms long intertrial interval, the new trial started. Importantly, participants did not see the other's action and were not provided any feedback about their failure/success on a given round, to eliminate the possibility of feedback-based learning.

Depending on the colour of the figure, trials could belong to one of the following four experimental conditions: blue, green, previously miscategorized and previously properly categorized ambiguous.

Participants were tested alone, in the same lighting conditions as before. The task started with eight introductory trials (two from each condition), that was followed by a message on the screen warning participants that they and the partner have already lost some points, so, they should try harder to be as fast as possible. The aim of this warning was to prompt participants to improve their performance (e.g. via considering the other's perspective). Following this, participants received two blocks of 32 test trials, with each block containing eight trials from each condition, and each shape appearing four times within each condition. To avoid that participants make errors because of the switch in the label-location associations, the side of the gates (BLUE left-GREEN right or GREEN left-BLUE right) corresponded to the side of the boxes in the anticipatory looking task and was fixed within participants.

The order of the trials within the blocks was pseudorandomized, so that there were no more than more than four consecutive trials with the same target gate location, condition type (previously properly or miscategorized), colour category (green or blue) and no more than three consecutive trials with the same colour and item (i.e. shape).

2.2.1.4. Explicit perspective-taking task

Stimuli and apparatus

Stimuli consisted of the animal pictures and the two 'boxes' of the anticipatory looking task and the geometric figures of the implicit transfer task. The apparatus and the settings were the same as before. The primary focus of interest was how participants categorize the geometric figures from the other's perspective (e.g. whether they would categorize the ambiguous green as blue, if the partner did so previously, in the anticipatory looking task). While categorizing the animal pictures in a way that would reflect the other's perspective could be done by simply recalling how their partner acted upon seeing a specific picture frame colour in the anticipatory task, this cannot apply to the shapes, as they have never seen the partner performing an action on these. Hence, one can safely assume that such a behaviour, that is, categorizing the shapes with the previously miscategorized ambiguous colour in a rule-incongruent manner, but in line with the partner's perspective, at least on some trials would reflect that they have updated the other agent's mental state in the anticipatory task upon encountering unexpected behaviour of the partner, and that they have encoded how she views a particular colour in a manner that is accessible to conscious decisions and generalizes beyond the actually observed stimuli.

Procedure

Participants were presented with four blocks of 32 trials, in two they had to perform colour-based categorization from their own perspective, and in the other two from the partner's perspective ('as their partner would categorize' the given item). They received written instructions before each block. The trial structure was similar to that of the 'implicit transfer task': each trial started with the presentation of a central fixation cross for 1000 ms, which was followed by the appearance of the two 'boxes' of the anticipatory task at the same location as the 'gates' before in the implicit transfer task.

1000 ms later a geometric figure or an animal picture appeared in the middle of the screen and stayed there until the participant clicked on it or for a maximum of 3000 ms. The presentation of the item was time-locked to the appearance of the cursor, the position of which was fixed at the bottom of the screen (at equal distance from the 'boxes'). After participants clicked on it, the geometric figure/animal picture disappeared and participants had a maximum of 2000 ms to perform the categorization. Selection of the box was followed by the appearance of the geometric figure/animal picture inside 300 ms after this, the trial ended and, following a 1000 ms long intertrial interval, the new trial started.

Stimulus presentation was blocked, with the geometric figures being presented first and the animal pictures second within each pair of blocks, to avoid that participants base the categorization of the geometric figures on the memory-traces about the partner's actions (how she categorized the pictures in the anticipatory task). Trials could belong to one of four conditions (blue, green, previously miscategorized and properly categorized ambiguous). Each block contained eight trials from each condition, with items counterbalanced within the conditions. The order of the trials was pseudorandomized using the same constraints that were used in the implicit transfer task before. The order of the block-pairs was counterbalanced across participants with half of them starting with the self-, the other half with the other-perspective blocks.

2.2.1.5. Post-test questionnaire

After the explicit perspective-taking task, participants received a short questionnaire, asking them 1) whether they have noticed anything peculiar in their partner's behaviour and 2) whether and to what extent did believe that they were interacting with another human (indicating their answer on a 5-point Likert scale). If their answer was 'yes' to the first question they were asked to specify what was peculiar about the other's behaviour and to reason about the potential causes (in writing).

Following the questionnaire, participants were extensively debriefed, revealing them that they were playing with a computer; none of them expressed any concern about this manipulation. The whole experiment lasted for approximately 80 minutes.

2.2.1.6. Data analysis

Anticipatory looking task

Eyetracking analyses were based on the raw gaze data of the participants. Gaze data were recorded separately for the two eyes but were averaged for each sample to obtain a more reliable measurement. Anticipatory looking was analysed in two time windows (segments) for each trial: in the 2000 ms long period after the onset of the target word until the beginning of the jitter (labelled as: anticipatory period 1), to capture potential early effects of the partner's perspective on participants' looking behaviour, and in the 1800 ms long period after the onset of the 'partner's' picture selection until the lighting up of the target box, i.e. right before the partner's action (labelled as: anticipatory period 2). Since this latter time window is the one in which expectations regarding the other agent's actions should exert the strongest effect, we focus on this period in the main text and discuss results from the first anticipatory period in detail only in the Supplementary Materials.

Segments for which more than 50% of the datapoints were missing and participants who had >30% missing segments were excluded from the further analyses. For each trial two rectangular areas of interest (AOIs) were defined, corresponding to the target box (the one that lit up in the given trial), and the incorrect box. Importantly, on the 'miscategorized' test trials the 'target AOI' was the box that was incorrect according to the task rules (e.g. greenish go to the green box) but corresponded to the 'partner's' updated mental state ('she thinks this greenish is blue') and hence her future actions.

Our main analyses focused on the *proportion of looking* at the target box. Proportion of looking was calculated for each trial and time window, separately, and was defined as the time spent with looking at the target box/time spent with looking at the two boxes altogether. If the participant did not look at any of the boxes in the given time window, the value was set as missing ('noanticipation' trials). In addition, we also analysed at which box the participant *looked first* in the respective time window, with no restriction set on the minimum length of the look. First look at the target box was coded as 1, at the incorrect box as 0. The results of these additional analyses are mentioned only briefly in the main text, with the details presented in the Supplementary Materials.

Data from the familiarization phase was analysed by Wilcoxon Signed Rank Tests, averaging the proportion of looking for all ambiguous and unambiguous trials (separately), to obtain a reliable measurement of anticipation, signalling that participants have formed initial expectations about the other's actions in this phase. Participants for whom the proportion of anticipatory looking in familiarization was lower than chance (0.5) in the ambiguous trials, were excluded from the further analyses.

Anticipatory looking in the test phase was analysed with Generalized Estimation Equations (GEE) clustered on individuals, assuming an exchangeable working correlation matrix and using maximum likelihood estimation method with robust estimator. In all analyses the default link function was used. By applying such a method we could treat (i) the proportion of looking data as normally distributed and the first look data as binomial and (ii) were able to include all participants in the analyses that would not have been possible in case of repeated-measures ANOVA, due to missing data. In all analyses, trial, condition and block served as within-subject variables and condition and block as predictors. Importantly, as the task-rule (how the pictures *should* be categorized) remained valid throughout the task and could be expected to exert a strong influence on anticipatory looking, akin to the 'pull of the real', demonstrated in other false-belief anticipatory looking paradigms (see e.g. Schneider et al., 2014), we expected an altercentric bias to emerge for the miscategorized colour, but not a full switch towards the belief-consistent (but rule-incongruent) target location. Therefore, in the test phase, we compared the proportion of looking to the target box on the miscategorized ambiguous trials to the corresponding control condition, specifically to the proportion of looking to the incorrect box on the properly categorized ambiguous colour trials. Pairwise comparisons were run on the estimated marginal means of the two conditions, for the four blocks, using Bonferroni-correction to adjust for multiple comparisons. For all post-hoc tests, the adjusted p-values are reported, along with z-scores and 95% Wald confidence intervals. For ease of reading figures display the raw means. Since our hypotheses concerned the difference between the two ambiguous colour conditions, analyses were run without including the other two conditions.

Reaction times were analysed with 2x4 repeated-measures ANOVAs (with condition and block as repeated-measures variables), and follow-up t-tests (all two-tailed, with p-values adjusted for multiple comparisons), using the correct responses only. Trials on which participants did not provide a response or clicked on the wrong box were excluded, as well as and reaction times more than three standard deviations from the condition means of each participant (calculated separately for each block). In case of Experiment 1, this meant the exclusion of 8.96% of the trials in the miscategorized and 7.92% trials in the properly categorized ambiguous condition. In case of Experiment 2, it meant the exclusion of 6.07% and 4.78% of the trials, respectively.

Implicit transfer and the explicit perspective-taking task

For both of these tasks the primary dependent measure was the hit rate on the previously miscategorized ambiguous trials (calculated separately for the four blocks in case of the explicit

perspective-taking and by averaging across all 16 trials in case of the implicit transfer task[6]), with comparisons made to the hit rate on the previously properly categorized ambiguous colour trials. Importantly, in both tasks passing/categorizing the presented item according to its actual colour (self-perspective) was coded as 1, while taking the other's perspective was coded as 0, therefore taking into account the other's (updated) perspective should be reflected in lower (and not higher) scores. Due to the non-normal distribution of the data all analyses on hit rates were run using nonparametric methods. All tests were two-tailed with significance level was set at $p<0.05$. To gain a more fine-grained picture of the processes underlying the participants' decisions, in case of the implicit transfer and the explicit perspective-taking task, we also analysed latencies (defined as the time it took for participants to click on the presented item from the appearance of the cursor)[7]. Latency data was analysed with paired-samples t-tests and Wilcoxon Signed Rank Tests (in case of non-normal distribution), excluding data more than three standard deviations from the condition means of each participant (calculated separately for each block and condition). For the implicit transfer task, this meant the exclusion of 7.29% and 5.41% of the data in the previously miscategorized and 6.88% and 6.46% of the data for the previously properly categorized condition, in Experiment 1 and Experiment 2, respectively. For the self-perspective trials of the explicit transfer task, these numbers were, 4.91%-6.70% and 4.91%-5.36 for the previously miscategorized figures and pictures and 4.91%-9.82% and 7.59%-7.59% for the previously properly categorized figures and pictures (for Experiment 1 and Experiment 2, respectively).

Analyses were first run on the whole sample for all three tasks. Then, we split the sample into two subgroups, on the basis of how participants performed on the other perspective trials of the explicit perspective-taking task. The 'update' subgroup consisted of participants whose explicit responses provided evidence that they had realized that the other agent represents the miscategorized (but not the other ambiguous) colour in a different way as they do (operationalized as: previously properly categorized ambiguous colour trials hit rate – miscategorized trials hit rate > 0). The 'noupdate' subgroup consisted of participants for whom there was no clear evidence for such a recomputation in their responses in the explicit perspective-taking task (operationalized as: previously properly categorized ambiguous – miscategorized trials hit rate =<0). Next, we reran the main analyses separately for the two subgroups.

[6] Introductory trials of the implicit transfer task were not included in the analyses.
[7] Since the items disappeared, this was the time window when participants actually had to make their decisions.

2.2.2 Results

2.2.2.1. Anticipatory looking task: looking behaviour

Familiarization phase

Analyses of the familiarization phase revealed that the proportion of looking towards the target box was significantly higher than chance (0.5), for both the ambiguous and the unambiguous trials (ambiguous: Z=-4.77, p < .001, r=.870; unambiguous: Z=-4.10, p < .001, r=.748), indicating that, by the time the test phase started, participants developed reliable expectations regarding where the upcoming events will take place (see **Table 2.1**)**.** The proportion of looking was significantly higher for the later miscategorized then for the later properly categorized ambiguous trials (Z=-2.95, p = .003, r=.538).

Table 2.1. *The mean proportion of looking towards the target box (correct anticipation) in the familiarization phase of Experiment 1 on the ambiguous and unambiguous colour trials, and separately for the two ambiguous trials, in the second anticipatory period (SD).*

	UNAMBIGUOUS	AMBIGUOUS	MISCATAMB	PROPCATAMB
	0.82 (0.24)	0.87 (0.14)	0.91 (0.15)	0.83 (0.14)

Note: MISCATAMB denotes the ambiguous colour that was later miscategorized by the partner (in the test phase), PROPCATAMB is the other ambiguous colour. *N*s vary due to missing data (no valid trials). For the MISCATAMB trials N=28.

Test phase

Figure 2.2a depicts the mean proportion of looking to the correct box in the miscategorized and to the incorrect box in the properly categorized ambiguous condition in the four blocks of the test phase. Higher anticipation in the miscategorized condition would signal that participants have updated how the partner encodes one of the ambiguous colours. The analysis yielded a significant main effect of condition (Wald χ^2=19.80, df=1, p< .001) and block (Wald χ^2=18.37, df=3, p< .001). Importantly, there was also a significant condition x block interaction (Wald χ^2=36.63, df=3, p< .001). While the proportion of looking towards the incorrect box in the properly categorized ambiguous condition tended to decrease with time, the proportion of looking towards the target box (that was incorrect from first-person but correct from third-person perspective) in the miscategorized condition increased sharply, indicating a revision of the original expectations. Pairwise comparisons revealed that the difference between the two conditions was significant from the second block on (miscategorized>properly categorized: block2 – z=4.38, M_{diff}= 0.27, Wald 95% CI [0.15-0.39], p_{adj}< .001; block3 – z=3.56, M_{diff}=

0.26, Wald 95% CI [0.12-0.41], p_{adj}< .001; block4 – z=5.41, M_{diff}= 0.37, Wald 95% CI [0.24-0.50], p_{adj}< .001). Analysis of the first look data indicated a similar pattern (see: Supplementary Materials, **Figure S2.2**). With respect to the proportion of looking in the first anticipatory period, the difference between the two conditions emerged only by the last block (see: Supplementary Materials, **Figure S2.4**).

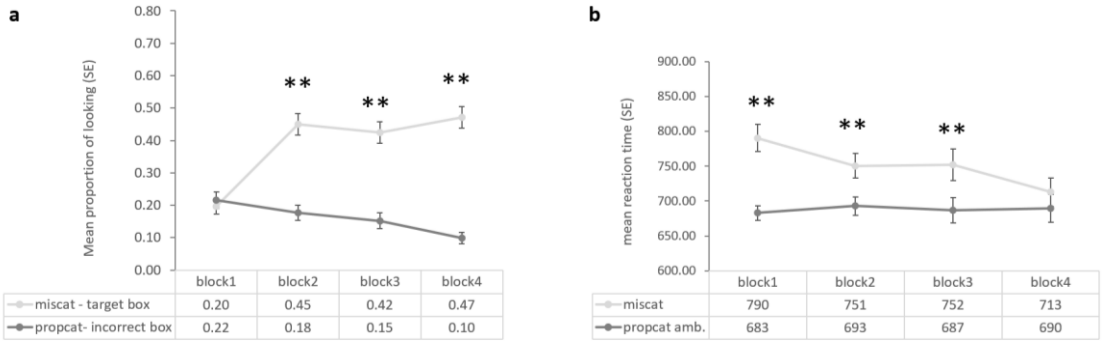


Figure 2.2. Changes in the (a) proportion of looking towards the target box in the miscategorized condition (miscat: light grey line) versus the incorrect box in the properly categorized ambiguous condition (propcat: dark grey line) the 'partner's' box selection and (b) mean reaction times in the miscategorized (light grey line) and properly categorized ambiguous (dark grey line) conditions per block during the test phase of the anticipatory looking task in Experiment 1. The proportion of looking figure displays the raw data. Proportion of looking analyses were run on the estimated marginal means. Error bars represent SE.+: p_{adj}<0.1; *: p_{adj}<0.05; **: p_{adj}<0.01.

2.2.2.2. Anticipatory looking task: behavioural measures

With respect to behavioural measures, hit rate was at ceiling in both the familiarization and the test phase of the task (>0.98 for all blocks in both conditions). Reaction times results are presented on **Figure 2.2b**. Analysis of reaction time data yielded a significant main effect of condition, F(1, 29)=33.30, p< .001, η_p^2= .535, and block, F(3, 87)=3.09, p= .031, η_p^2= .096, as well as a significant condition x block interaction, F(3, 87)=8.22, p< .001, η_p^2= .221. While RT did not change substantially over time in the properly categorized ambiguous condition (indicated by the fact that none of the pairwise comparisons between blocks was significant: all ts<1.89, all ps >.379) it did so in the miscategorized condition. Participants were much slower on these trials in the first block, than in the subsequent ones, with the difference being significant compared to the second (t(29) = 2.97, p_{adj}= .018, d=0.542) and the fourth block (t(29) = 5.40, p_{adj}< .001, d=0.986) but not in comparison to the third block, after adjusting for multiple comparisons (t(29) = 2.04, p_{adj}= .150, d=0.373). Reaction times also decreased substantially after the third block with a marginally significant difference between the fourth and the previous two blocks (block4 vs block2: t(29) = 2.54, p_{adj}= .051, d=0.463; block4 vs block: t(29) = 2.54, p_{adj}= .051, d=0.465). Pairwise comparisons revealed that the difference between the miscategorized and properly

categorized ambiguous condition was significant throughout the first three blocks (block1: $t(29) = 7.07$, $p_{adj} < .001$, d=1.29; block2: $t(29) = 4.51$, $p_{adj} < .001$, d=0.823; block3: $t(25) = 4.06$, $p_{adj} < .001$, d=0.742) but disappeared by the last block ($t(29) = 1.40$, $p_{adj} = .688$, d=0.256), suggesting that by then participants fully adapted to the partner's initially surprising behaviour.

2.2.2.3 Implicit transfer task

Wilcoxon Signed Rank Test revealed that the hit rate was significantly lower on the previously miscategorized than on the previously properly categorized ambiguous colour trials (Z=-2.46, p= .014, r=0.449), indicating that participants took into account how the other sees the colour she miscategorized earlier in the anticipatory looking task and generalized this knowledge to a new task. However, the difference was rather small (M_{diff}=0.11, SD_{diff}=0.27, see **Figure 2.3a**). Inspection of the individual data showed that there were only 11 participants who provided evidence for such an implicit transfer, by making 'errors' (i.e. passing the geometric figures congruent with the other's perspective) on the previously miscategorized but not on the properly categorized ambiguous trials. Out of these 11 only 2 passed the geometric shapes according to the partner's (updated) perspective on more than 50% of the trials, the remaining 9 participants took into account the other's perspective at least once (on 6%-50% of the trials). With respect to response time, there was no difference between the two conditions ($t(29)$= -0.469, p= .643, d=0.090; M_{diff}=-3 ms, SD_{diff}=40, see **Figure 2.3b**).

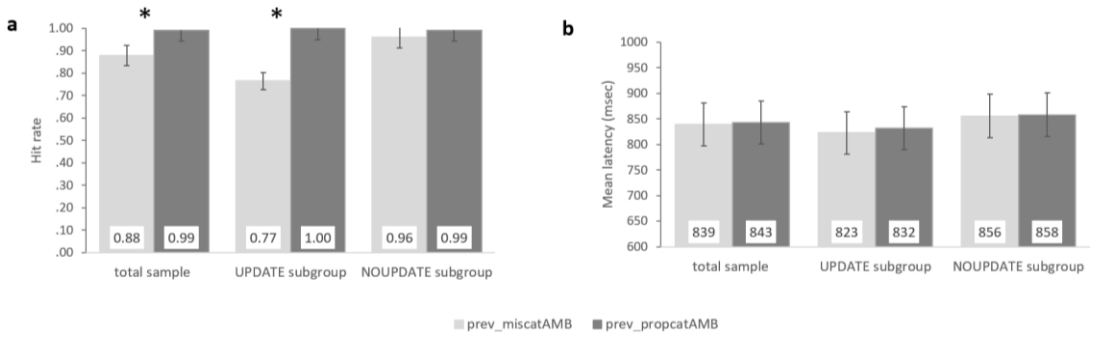


Figure 2.3. (a) Mean hit rate and (b) latency on the previously miscategorized (prev_miscatAMB) and previously properly categorized ambiguous trials (prev_propcatAMB) of the implicit transfer task, for the whole sample (left panels) and for the UPDATE (middle panels) and NOUPDATE subgroups (right panels), in Experiment 1, created on the basis of participants' performance on the explicit perspective-taking task. Passing the geometric figures according to the 'partner's' updated perspective was coded as 0, hence was associated with lower hit rates on the previously miscategorized colour trials. Error bars represent SE. +: p<0.1; *: p<0.05; **: p<0.01.

2.2.2.4 Explicit perspective-taking task

We could analyse the data from 28 participants (15 males, age: M_{age}= 23.64, SD_{age}=4.21). Data from two additional participants could not be included due to technical problems that arose during the task. **Figure 2.4a** and **2.4b** displays the mean hit rates and the response latencies for the previously miscategorized and previously properly categorized ambiguous conditions in the self- and the other-perspective blocks, separately for the geometric figures of the implicit transfer and the animal pictures of the anticipatory looking task. As can be seen from the figure, hit rates were numerically lower for the previously miscategorized than for the previously properly categorized ambiguous colour trials in the other perspective blocks, both for the geometric figures and the animal pictures, although performance was close to ceiling on the same trials in the self-perspective blocks. Importantly, the hit rate was lower in the miscategorized condition when participants had to make judgements from the other's compared to their own perspective (miscategorized condition hit rate - figures: self > other, Z=-3.30, p< .001, r=0.624; pictures: self > other, Z=-3.12, p= .002, r=0.590). Such results indicate that participants were aware of the fact that the other perceived the previously miscategorized colour in a different way as they themselves did and could take this into account to some extent when performing the categorization from her perspective. However, the difference between the two types of ambiguous trials was not significant in either of the two other perspective blocks (figures: Z=-1.10, p= .272, r=0.208; pictures: Z=-1.51, p= .132, r=0.285) due to the fact that participants made errors on the previously properly categorized ambiguous trials as well (i.e. categorized these items in a rule-incongruent manner). This implies that they were not entirely successful in identifying or recalling which of the two ambiguous colours their partner perceived in a different way previously.

A closer look at the data revealed that the individual variation in adjusting to the other's (updated) perspective was relatively high: 7 participants categorized the geometric figures according to the other's (updated) perspective on all eight miscategorized trials (but did not make any mistake on the properly categorized ambiguous trials), 14 did not show any sign of correct perspective-taking, and 7 took it into account on 1 to 7 trials. Spearman' correlation indicated a strong relationship between the hit rate of the figure and picture trials (other-perspective blocks - previously miscategorized condition: r_s=.861, p<.001). With respect to latencies, there was no significant difference between the two conditions: it took equally long (or even somewhat shorter time) for participants to make their decision on the previously miscategorized colour trials than to categorize the items on the previously properly categorized ambiguous colour trials, both when taking the other's perspective (figure: t(28)= -0.493, p= .626, d=0.009; picture: t(28)= 0.034, p= .974, d=0.006) and when performing categorization from their own (self-perspective - figure: t(28)= -0.097, p= .924, d=0.018 ; picture: t(28)= -2.041, p= .051, d=0.386, previously properly categorized > miscategorized).

Upon receiving the post-test questionnaire, 25 participants reported noticing the miscategorization and 18 gave mentalistic accounts for it (such as the other saw the colour in a different way or was presented with a different colour on the miscategorization trials).

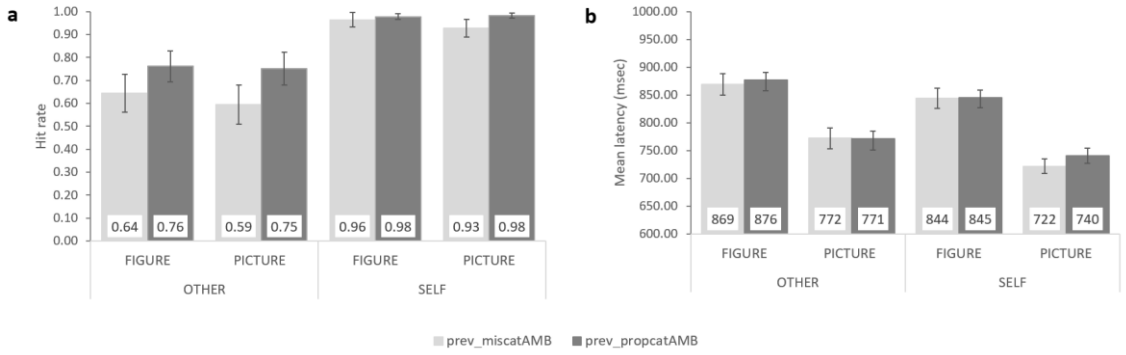


Figure 2.4. (a) Mean hit rate) and (b) latency on the previously miscategorized (prev_miscatAMB) and previously properly categorized ambiguous colour trials (prev_propcatAMB) of the explicit perspective-taking task, in Experiment 1. FIGURE denotes the geometric figures used in the implicit transfer task, PICTURE refers to the animal pictures used in the anticipatory looking task. OTHER vs SELF: blocks in which participants had to categorize the items from their 'partner's' versus their own perspective. Lower hit rate on the previously miscategorized trials in the OTHER blocks reflect perspective-taking. Latency is the time elapsed from the appearance of the cursor until the participant clicked on the figure. Error bars represent SE.
+: p<0.1; *: p<0.05; **: p<0.01

2.2.2.5 Subgroup analyses based on the performance on the explicit task

Splitting the sample along participants' performance on the other perspective trials (i.e. along previously properly categorized – previously miscategorized figures' hit rate, with the difference score = 0 used as a cut-off point) yielded 12 participants in the 'update' (difference score > 0; 5 males, age: M_{age}= 22.58, SD_{age}=3.50) and 16 in the 'noupdate' subgroup (difference score≤ 0; 9 males, age: M_{age}= 24.44, SD_{age}=4.61). The two groups did not differ in age ($t(26)$=-1.16, p= .256, d=0.454) or gender (Fisher's Exact Test: p= .704).

2.2.2.5.1 Update subgroup

Anticipatory looking task: proportion of looking (test phase)

Figure 2.5a presents the mean proportion of looking for the four blocks of the test phase in the two conditions for the 'update' subgroup. Analysis revealed a significant main effect of condition (Wald χ^2=28.28, df=1, p< .001) and block (Wald χ^2=15.31, df=3, p= .002) as well as a significant condition x block interaction (Wald χ^2=35.73, df=3, p< .001). Pairwise comparisons revealed significant difference between the two conditions from the second block on (miscategorized>properly categorized ambiguous: block2 – z=5.10, M_{diff}= 0.46, Wald 95% CI [0.28-0.63], p_{adj}< .001; block3 – z=4.42, M_{diff}= 0.50, Wald 95% CI [0.28-0.72], p_{adj}< .001; block4 – z=6.24, M_{diff}= 0.61, Wald 95% CI [0.42-0.80], p_{adj}< .001), indicating that by then participants updated their partner's mental state and revised their expectations regarding how the partner would act upon seeing the specific, miscategorized colour. In the first anticipatory period, the difference between the two conditions emerged only by the last block (see: Supplementary Materials **Figure S2.3a**).

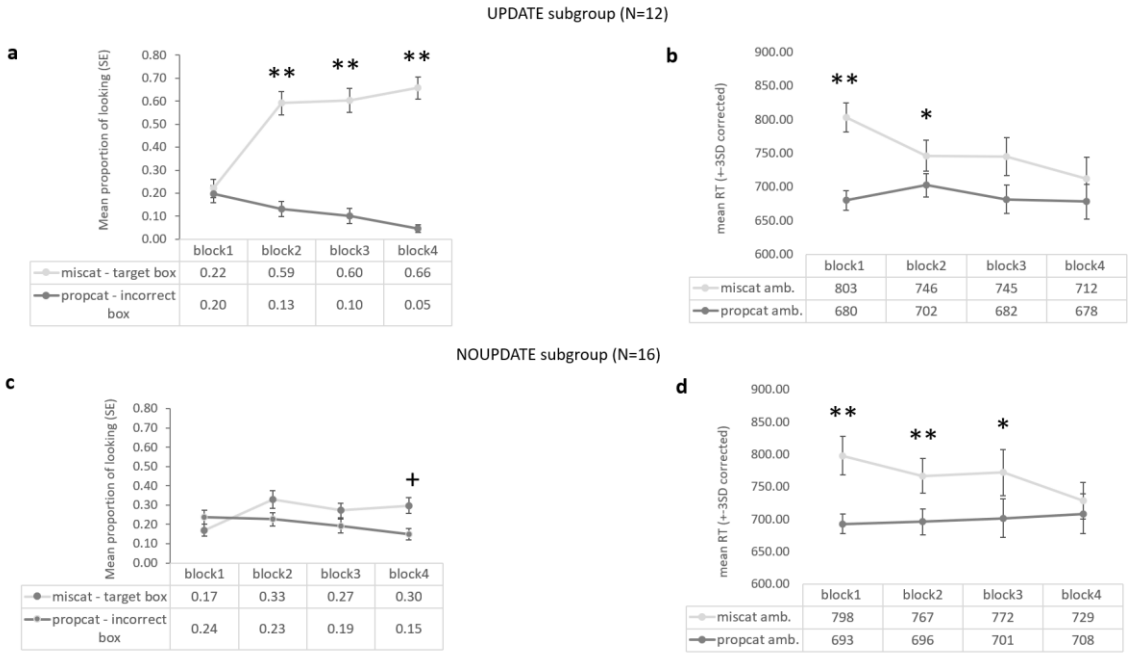


Figure 2.5. Changes in the proportion of looking towards the target box in the miscategorized condition (miscat: light grey line) versus the incorrect box in the properly categorized ambiguous condition (propcat: dark grey line), prior the partner's box selection, and (b) in mean reaction times in the miscategorized (light grey line) and properly categorized ambiguous (dark grey line) conditions in the (a, b) UPDATE (upper panel) and (c, d) NOUPDATE (lower panel) subgroups of Experiment1. The subgroups were created on the basis of participants' performance on the other-perspective trials of explicit perspective-taking task. The figure displays the raw data. Error bars represent SE. +: p_{adj}<0.1; *: p_{adj}<0.05; **: p_{adj}<0.01.

Anticipatory looking task: reaction time

Analysis of the reaction times revealed a significant main effect of condition, $F(1,11)=27.87$, $p< .001$, $\eta_p^2= .717$ but no significant main effect of block, $F(2.00,21.98)=2.35$, $p= .119$, $\eta_p^2= .176$. There was, however, a significant condition x block interaction, $F(3,33)=5.03$, $p= .006$, $\eta_p^2= .314$. As can be seen on **Figure 2.5b** 'update' participants were much slower on the first few trials of the miscategorized condition than on the subsequent ones, with the reaction times dropping markedly following the first and the third block. After adjusting for multiple comparisons, the difference turned out to be significant for the first versus the last but not for the other blocks (block1 vs block4: $t(11) = 3.12$, $p_{adj}= .030$, $d=0.709$; for the other comparisons: all ts<2.45, and ps > .096), with the other comparisons not being significant either (all ts<1.62, and ps > .134). There was no such change over the course of trials in the properly categorized ambiguous condition (all ts<1.56, all ps > .148). Pairwise comparisons revealed that the difference between the two conditions was significant for the first ($t(11) = 7.79$, $p_{adj}< .001$, $d=2.25$) and the second block ($t(11)= 3.21$, $p_{adj}= .032$, $d=0.926$) but became nonsignificant by the last two blocks, after adjusting for multiple comparisons (block3: $t(11)= 2.39$, $p_{adj} = .144$, $d=0.690$; block4: $t(11)= 1.59$, $p_{adj} = .556$, $d=0.460$).

Implicit transfer task

The analysis revealed a significantly lower hit rate for the previously miscategorized than for the previously properly categorized ambiguous colour trials (Z=-2.38, $p= .017$, $r=0.688$), indicating that participants took into account how the other sees the previously miscategorized the colour and used what they have learnt about the other in this new task. As can be seen from **Figure 2.3a** (middle panel) the difference was larger than for the whole sample, reflecting the fact that six of the nine participants who demonstrated signs of perspective-taking in the implicit transfer task belonged to the update subgroup. There was, however, no difference between the two conditions in how much time it took participants to make their decisions ($t(11)= -0.703$, $p= .497$, $d=0.203$).

Explicit perspective-taking task

With respect to the explicit perspective-taking task, despite participants being slower in categorizing the previously miscategorized than the previously properly categorized animal pictures from the partner's perspective, this difference was not significant ($t(11)= 1.46$, $p= .172$, $d=0.422$). There was no

difference in how fast they categorized the two types of figures on the other-perspective trials ($t(11)$= -0.37, p= .717, d=0.107). Analyses revealed no significant difference regarding how fast participants could categorize the previously properly categorized and miscategorized items from their own perspective either (figures: $t(11)$= -0.441, p= .668, d=0.127; pictures: $t(11)$= -0.305, p= .766, d=0.088), suggesting no interference from the other's perspective. However, participants made somewhat more errors when they had to categorize animal pictures with the previously miscategorized colour from their own perspective, compared to when they had to categorize pictures the frame of which had the previously properly categorized colour (Z=-1.76, p= .078, r=0.509) (see **Figure 2.6a** and **2.6b**)**.** No such difference was present for the figures (Z=-1.00, p= .317, r=0.289). In the post-test questionnaire, 9 participants could specify which colour was miscategorized/misperceived by the other and 8 gave mentalistic accounts for the observed change.

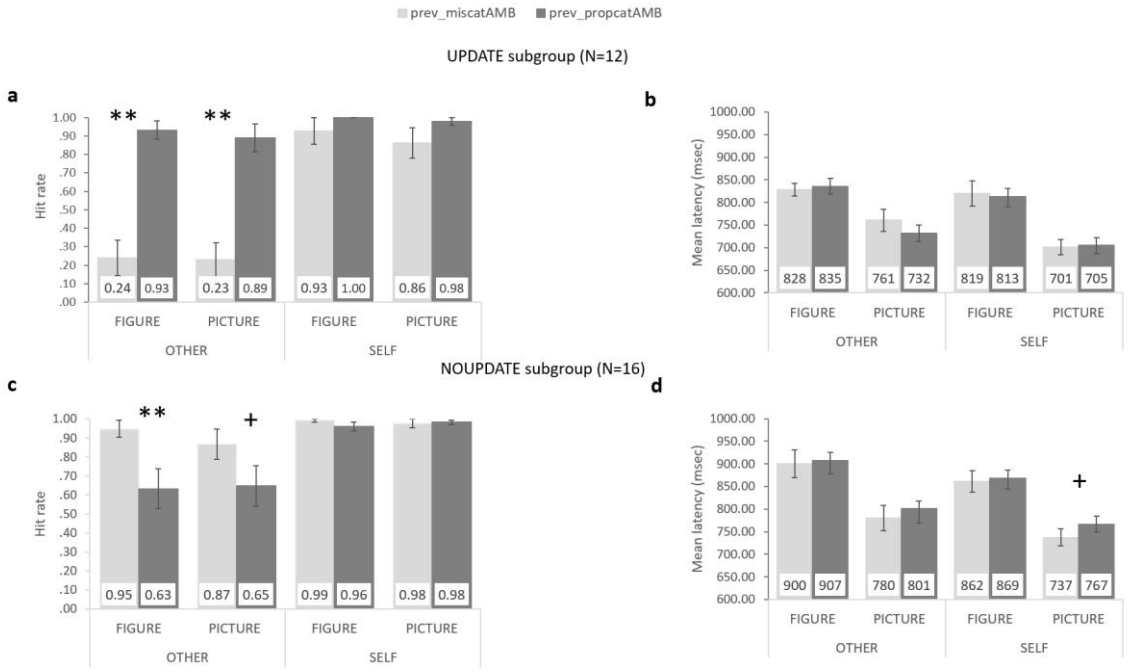


Figure 2.6. Mean hit rate (a, c) and latency (b, d) on the previously miscategorized (prev_miscatAMB) and previously properly categorized ambiguous trials (prev_propcatAMB) of the explicit perspective-taking task, in the 'update' (upper panel) and 'noupdate' (lower panel) subgroups of Experiment 1. FIGURE denotes the geometric figures used in the implicit transfer task, PICTURE refers to the animal pictures used in the anticipatory looking task. OTHER vs SELF: blocks in which participants had to categorize the items from their 'partner's' versus their own perspective. Lower hit rate on the previously miscategorized trials in the OTHER blocks reflect perspective-taking. Error bars represent SE. Note that the figure also displays the hit rate on the OTHER trials, to provide a full picture of the data pattern. However, as we used the hit rate on the previously miscategorized and properly categorized trials to create the two subgroups, we do not compare these trials statistically. +: p<0.1; *: p<0.05; **: p<0.01

2.2.2.5.2 Noupdate subgroup

Anticipatory looking task: proportion of looking (test phase)

Figure 2.5c presents the mean proportion of looking for the four blocks of the test phase in the two critical conditions for the 'noupdate' subgroup. Analysis of the proportion of looking data yielded no significant main effect of the condition (Wald χ^2=2.52, df=1, p= .112). There was, however, a significant main effect of block (Wald χ^2=9.89, df=3, p= .019), and a significant condition x block interaction (Wald χ^2=15.54, df=3, p= .001): as the experiment unfolded, participants tended to look more towards the target box on the miscategorized trials and somewhat less towards the incorrect box on the properly categorized ambiguous trials, with pairwise comparisons indicating a tendency level difference between the two conditions by the last block (miscategorized> properly categorized: z=2.48, M_{diff}= 0.14, Wald 95% CI [0.03-0.26], p_{adj}= .064), implying that, by the end of the experiment, they started to revise their original expectations. No such effect was present in the first anticipatory period (see Supplementary Materials **Figure S2.3c**).

These results were not due to a generally lower level of anticipation in this subgroup: the number of trials where participants did not look at any of the boxes (proportion of 'noanticipation' trials) did not differ significantly between the two subgroups (miscategorized: Z=-0.40, p= .698, r=0.073; properly categorized ambiguous: Z=-0.19, p= .873, r=0.035, see **Table S2.3** for details).

Anticipatory looking task: reaction time

Analysis of reaction time data yielded no significant main effect of block, F(1.92,28.77)=0.856, p= .431, η_p^2= .054, but a significant main effect of condition, F(1,15)=13.58, p= .002, η_p^2= .475, as well as a significant condition x block interaction, F(3,45)=3.94, p= .014, η_p^2= .208. While there was no substantial change over time in participants' reaction times in the properly categorized ambiguous condition (with none of the pairwise comparisons between blocks being significant: all ts<0.67, all ps >.514) there was a change in the miscategorized condition. Just like 'update' participants members of the 'noupdate' subgroup were slower on the first few miscategorized trials, than on the subsequent ones. As can be seen on **Figure 2.5d** the decrease was, however, much smaller in this subgroup after the first block and became pronounced only by the last block. The difference was significant only for the first versus the last block (t(15) = 4.74, p_{adj}< .001, d=1.186), but not for the other blocks (all ts<2.27, and p_{adj}s > .117) or any of the other comparisons, after adjusting for multiple comparisons (all ts<2.01, and p_{adj} s > .189), indicating a rather slow adaptation to the change in the other's behaviour. The

difference between the two conditions was significant in the first three blocks (block1: $t(15) = 4.52$, $p_{adj} < .001$, d=1.13; block2: $t(15) = 3.27$, $p_{adj} = .020$, d=0.817; block3: $t(15) = 3.06$, $p_{adj} = .032$, d=0.764) but disappeared by the last block ($t(29) = 0.77$, $p_{adj} = 1.00$, d=0.192).

Implicit transfer task

Regarding the implicit transfer task, hit rate was close to ceiling on both the previously miscategorized and previously properly categorized ambiguous colour trials. The analysis indicated no significant difference between the two conditions (Z=-0.92, p= .357, r=0.238), revealing that, unlike members of the 'update' subgroup, 'noupdate' participants did not take into account that other's different perspective in the task. There was no difference between the two conditions in terms of the latencies either ($t(15) = -0.25$, p= .803, d=0.063).

Explicit perspective-taking task

Just like members of the 'update' subgroup, 'noupdate' participants were equally fast in categorizing the previously miscategorized and properly categorized ambiguous figures ($t(15) = -0.33$, p= .744, d=0.107) and somewhat slower when categorizing the previously properly categorized than the previously miscategorized animal pictures figures from their partner's perspective ($t(15) = -0.83$, p= .418, d=0.422). However, when categorizing the animal pictures from their own perspective, they tended to be slower ($t(15) = -2.40$, p= .030, d=0.601, M_{miscat}=737 ms, SD_{miscat}=75ms versus $M_{propcat}$=767 ms, $SD_{propcat}$=71ms) on the previously properly categorized compared to the previously miscategorized ambiguous colour trials, indicating a possible confusion regarding which of the two ambiguous colours was miscategorized previously by the partner. For the geometric figures, there was no such difference in the latencies ($t(15) = -0.35$, p= .734, d=0.087). The two conditions did not differ in terms of hit rate for either of the two types of items (animal pictures: Z=-0.447, p= .655, r=0.116; geometric figures: Z=-1.3, p= .194, r=0.376). Interestingly, despite their performance on the explicit perspective-taking task, 8 of the 'noupdate' participants gave a mentalistic explanation for the change in the partner's behaviour following the explicit task.

2.2.3 Discussion

Results of Experiment 1 provided evidence that participants spontaneously revised their expectations regarding their 'partner's' future behaviour and updated her perspective, upon seeing her categorizing one of the colours in a different way than before. In the anticipatory looking task correct anticipations started to emerge after the first few miscategorized trials, significantly exceeding what could be expected by random looking already by the second block. Such a change in anticipatory looking was accompanied by a fast behavioural adaptation, indicated by a marked drop in participants' reaction times, following an initial increase, that likely reflected surprise. Results of the explicit perspective-taking task, in specific the correct, rule-congruent categorization of the previously miscategorized colour on the self-perspective trials, indicate that the observed change in the anticipatory looking indeed reflected updating of the partner's mental state and not participants' own belief regarding what categories the colours belong to (i.e. revision of the task-rule).

In addition, the performance of roughly 40% of the participants on the explicit perspective-taking task provided evidence that such a revision reflected a correct mentalistic interpretation of the change in the other's behaviour. In particular, these participants' performance indicated that they realized that their partner encoded a particular colour in a different way than they themselves did. They were able to categorize the previously miscategorized items from their 'partner's' point of view when explicitly instructed to do so, and this was the case not only for the animal pictures, used in the anticipatory task, where it was possible to merely recall how the other acted before, but also for the geometric figures of the implicit transfer task where they could not rely on such a strategy.

Further analysis of the eyetracking and reaction times data revealed that the effects observed at the group level were in great part driven by the looking behaviour of the 'update' subgroup, although they emerged in the 'noupdate' subgroup as well, by the end of the task. However, note that the extent to which participants took into account the 'partner's' (updated) perspective spontaneously in the subsequent implicit transfer task was relatively low even in this, 'update' subgroup. Despite the fact that the information about the characteristics of the other's perception was likely consciously available for these participants (as evidenced by their performance on the explicit perspective-taking task), they rarely made use of it in the implicit transfer task and adjusted their behaviour to the other's perspective only on a few trials.

It is worth noting that besides the surprisingly low level of implicit transfer (i.e. generalization), correct anticipation also remained rather low in the total sample on the miscategorized trials, not exceeding 0.5 even by the end of the anticipatory looking task. Although striking at the first glance, these results may not be surprising given the characteristics of our task: to update the 'partner's mental state and

use the information about it later, participants did not only have to overcome a general egocentric bias (i.e. how they see the colour), but they also had to overwrite a presumably strong first-person colour-colour label association as well as an explicit categorization rule that remained valid throughout the task, on the majority of the trials. These two features might have made the revision of expectations and use of the acquired information about the other's perspective rather difficult in our experiment. To test this possibility, and to investigate how updating and transfer of the knowledge takes place if no such explicit categorization rule is present and no colour-label association is involved, we ran a second experiment in which the 'partner' had to make similarity judgements instead of categorical ones (e.g. had to put the item in the box that was judged as more similar in colour). Such a manipulation could not only reduce the task demands but could help participants to realize that the other's miscategorization reflects her subjective evaluation (i.e representation) of the stimuli that may change over time (more than a categorical decision).

2.3. Experiment 2

Experiment 2 differed from Experiment 1 in only one aspect: after selecting the picture, the partner had to put it into the box which was judged more similar in colour to the colour of the selected picture's frame. In specific, the partner had to put the picture where she 'thought it belongs to' (on the basis of the colour). This manipulation was expected to highlight the subjective nature of the partner's decisions, and, as a result, boost updating and subsequent consideration of the other's perspective.

2.3.1 Methods

2.3.1.1 Participants

37 students were recruited to the experiment via a student job agency and the university's research participation system (SONA systems). The data of 3 participants was excluded because they did not meet the inclusion criteria for the eyetracking analysis (had too many missing datapoints: n=1 or did not develop correct anticipation in the familiarization phase of the anticipatory looking task: n=2). Thus, the final sample consisted of 34 participants (age: M_{age}= 23.82, SD_{age}=4.68, 17 males). All of them were right-handed and had normal or corrected-to-normal vision. The study was approved by the

Hungarian Ethical Review Committee for Research in Psychology; participants signed informed consent prior to the experiment and received monetary compensation or gift vouchers for their participation (equivalent to approximately 7 Euros).

2.3.1.2 Stimuli, apparatus and procedure

Stimuli, apparatus and procedure used were the same as in Experiment 1, in all three tasks, with two crucial exceptions. First, instead of colour labels, blue and green coloured squares were presented on the 'boxes' (for the anticipatory looking and the explicit perspective-taking tasks) and on the 'gates', in the implicit transfer task. These two colours were different shades of blue/green compared to the ones used for the other stimuli and were clearly identifiable as blue/green just like the unambiguous colours used in the task. The RGB values of the blue and green square were 3-30-209 and 9-159-2, respectively (for an example see **Figure 2.7**). Second, in the anticipatory task, after the animal's selection, the 'partner' had to categorize the items based on how similar the colour of the picture frame was to that of the square on the boxes, according to their view, placing the picture into the box where she 'thought it belongs' (on the basis of the frame's colour).

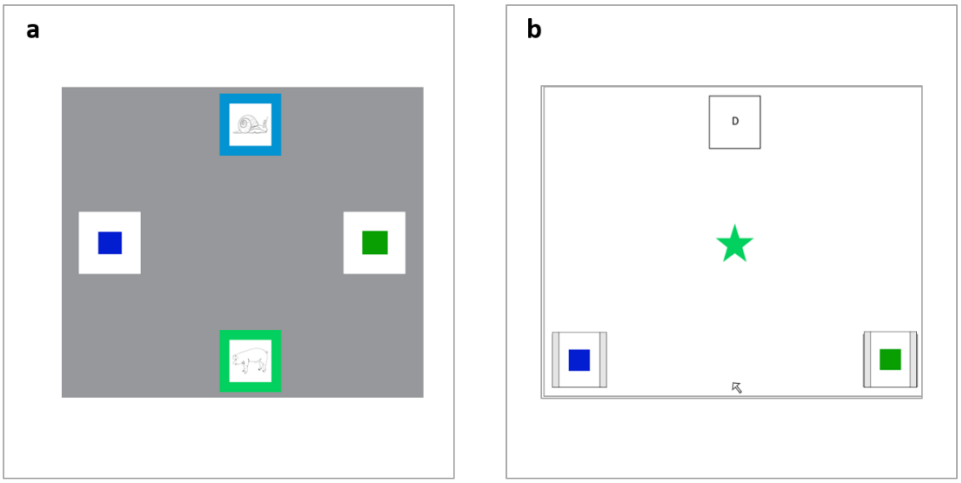


Figure 2.7. An example for the layout of the anticipatory looking task and the implicit transfer task in Experiment 2 (colour matching). The task was the same as in Experiment 1.

2.3.2 Results

2.3.2.1. Anticipatory looking task: looking behaviour

Familiarization phase

Analysis of the proportion of looking data in the familiarization phase indicated that participants developed a reliable anticipation towards the target locations by the beginning of the test phase, despite the lack of an explicit categorization rule, with the level of correct anticipation being even higher than in Experiment 1 (see **Table 2.2**). Wilcoxon Signed Rank Tests revealed that the proportion of looking towards the target box was significantly higher than chance (0.5) in the familiarization phase, on both the ambiguous and unambiguous trials (ambiguous: Z=-4.57, p < .001, r=.784; unambiguous: Z=-4.74, p < .001, r=.813). There was no significant difference between the two ambiguous conditions (Z=-0.18, p = .868, r=.031).

Table 2.2. *The mean proportion of looking towards the target box (correct anticipation) in the familiarization phase of Experiment 2 on the ambiguous and unambiguous colour trials, and separately for the two ambiguous trials, in the second anticipatory period (SD).*

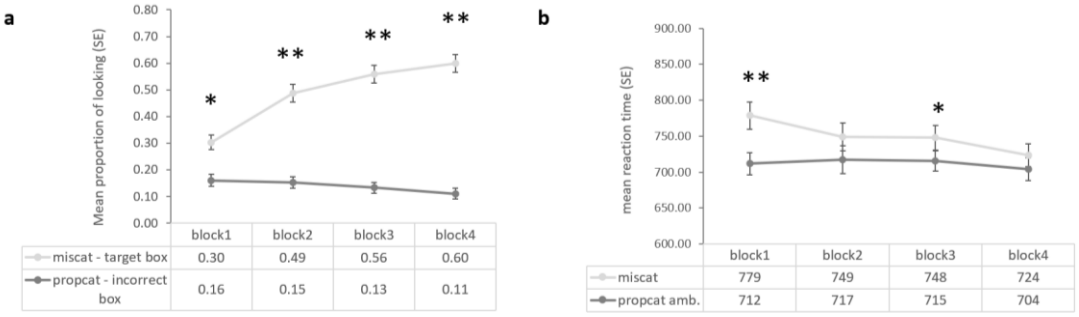
	UNAMBIGUOUS	AMBIGUOUS	MISCATAMB	PROPCATAMB
Experiment 2	0.88 (0.26)	0.84 (0.22)	0.85 (0.22)	0.84 (0.22)

Note: MISCATAMB denotes the ambiguous colour that was later miscategorized by the partner (in the test phase), PROPCATAMB is the other ambiguous colour. *N*s vary due to missing data (no valid trials). For the MISCATAMB trials N=33.

Test phase

Figure 2.8a presents the mean proportion of looking to the correct box in the miscategorized and to the incorrect box in the properly categorized ambiguous condition in the four blocks of the test phase, prior the partner's box selection. An analysis of the data yielded a significant main effect of condition (Wald χ^2=35.84, df=1, p< .001) and block (Wald χ^2=16.94, df=3, p= .001). There was also a significant condition x block interaction (Wald χ^2=40.01, df=3, p<.001), resulting from a marked increase in the proportion of looking towards the target box in the miscategorized condition after the first few trials. Pairwise comparisons revealed that the difference between the miscategorized and properly categorized ambiguous conditions was significant already from the first block (miscategorized>properly categorized: block1 – z=2.59, M_{diff}= 0.15, Wald 95% CI [0.04-0.26], p_{adj}= .040;

block2 – z=4.83, M_{diff}= 0.34, Wald 95% CI [0.20-0.47], p_{adj}< .001; block3 – z=6.54, M_{diff}= 0.43, Wald 95% CI [0.30-0.56], p_{adj}< .001; block4 – z=6.37, M_{diff}= 0.49, Wald 95% CI [0.34-0.64], p_{adj}< .001), suggesting that participants quickly updated the mental state attributed to the other and revised their expectations regarding his/her behaviour. Analysis of the first look data in the second anticipatory period and the proportion of looking data in the first anticipatory period indicated a similar pattern, with the difference being present from the second block on (see Supplementary Materials **Figure S2.2** and **S2.4**).



	block1	block2	block3	block4
miscat - target box	0.30	0.49	0.56	0.60
propcat - incorrect box	0.16	0.15	0.13	0.11

	block1	block2	block3	block4
miscat	779	749	748	724
propcat amb.	712	717	715	704

Figure 2.8. Changes in the (a) proportion of looking towards the target box in the miscategorized condition (miscat: light grey line) versus the incorrect box in the properly categorized ambiguous condition (propcat: dark grey line) the partner's box selection and (b) in the mean reaction times in the miscategorized (light grey line) and properly categorized ambiguous (dark grey line) conditions per block during the test phase of the anticipatory looking task in Experiment 2. The proportion of looking figure displays the raw data. Proportion of looking analyses were run on the estimated marginal means. Error bars represent SE.
+: p_{adj}<0.1; *: p_{adj}<0.05; **: p_{adj}<0.01

2.3.2.2. Anticipatory looking task: behavioural measures

With respect to behavioural measures, hit rate was at ceiling in both the familiarization and the test phase of the task (>0.99 for all blocks in both conditions). Analysis of the reaction time data revealed a significant main effect of condition, F(1,33)=32.85, p< .001, η_p^2= .499, and block, F(2.22,73.38)=3.43, p= .033, η_p^2= .094, as well as a significant condition x block interaction, F(3,99)=3.43, p= .020, η_p^2= .094, resulting from the fact that participants were slower on the first few miscategorized trials than on the subsequent ones, with a significant difference between the first versus the last block (t(33) = 3.50, p_{adj}= .003, d=0.560) and a tendency level difference between the first and the third block (t(33) = 2.24, p_{adj}= .096, d=0.384), but no significant difference between the first and the second block (t(33) = 1.92, p_{adj}= .189, d=0.330). Reaction times dropped again by the last block but the difference was only marginally significant compared to the previous (t(33) = 2.25, p_{adj}= .093, d=0.386) and not significant compared

to the second block, after adjusting for multiple comparisons ($t(33) = 1.87$, $p_{adj} = .210$, $d = 0.321$), implying a rather gradual behavioural adaptation (see **Figure 8b**). There was no such change over the course of trials in the properly categorized ambiguous condition (all ts<1.38, all ps > .177). Pairwise comparisons revealed that the difference between the two conditions was significant for the first ($t(33) = 6.54$, $p_{adj} < .001$, $d = 1.12$) and the third ($t(33) = 2.96$, $p_{adj} = .024$, $d = 0.507$) but not for the second block, after adjusting for multiple comparisons ($t(33) = 2.25$, $p_{adj} = .108$, $d = 0.398$). Importantly, the difference disappeared by the last block ($t(33) = 1.75$, $p_{adj} = = .356$, $d = 0.300$), indicating that by the end of the task participants could completely adapt to the change in the partner's behaviour.

2.3.2.3. Implicit transfer task

Wilcoxon Signed Rank Test revealed a significantly lower hit rate for the previously miscategorized than for the previously properly categorized ambiguous colour trials ($Z = -2.10$, $p = .044$, $r = 0.360$), showing that participants took into account their partner's different perspective when passing her the figures (see **Figure 2.9a**). Latency analyses indicated that it also took them significantly more time to make their decisions on these trials than on the previously properly categorized ambiguous colour trials ($t(33) = 2.59$, $p = .014$, $d = 0.443$; $M_{diff} = 32.75$ ms, $SD_{diff} = 73.84$, see **Figure 2.9b**). Detailed inspection of the individual data showed that there were 15 participants who made such 'errors' (reflecting the adoption of the other's perspective) on the previously miscategorized but not on the properly categorized ambiguous trials. Out of these 15, 9 participants passed the geometric figures according to their partner's perspective on more than 50% of the trials from the previously miscategorized colour condition, with an additional 6 doing so at least once during the task (on 6-50% of the trials).

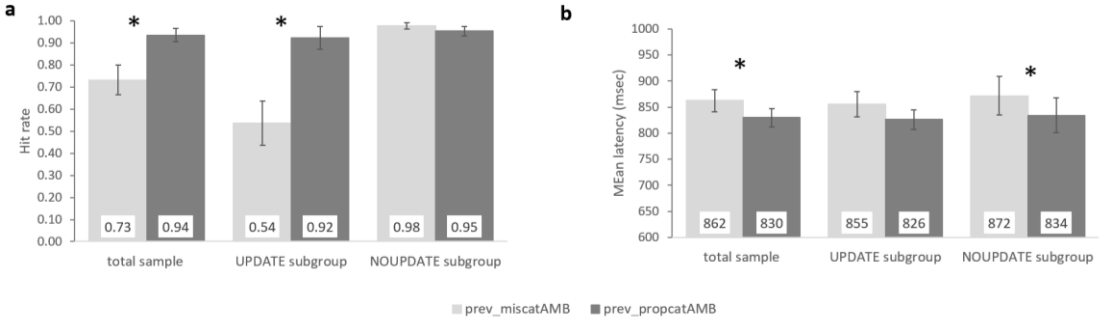


Figure 2.9. (a) Mean hit rate and (b) latency on the previously miscategorized (prev_miscatAMB) and previously properly categorized ambiguous trials (prev_propcatAMB) of the implicit transfer task, for the whole sample (left panels) and for the UPDATE (middle panels) and NOUPDATE subgroups (right panels), created on the basis of participants' performance on the explicit perspective-taking task. Passing the geometric figures according to the partner's updated perspective was coded as 0, hence was associated with lower hit rates on the previously miscategorized colour trials. Error bars represent SE. +: p<0.1; *: p<0.05; **: p<0.01.

2.3.2.4. Explicit perspective-taking task

Figure 2.10a and **2.10b** shows the mean hit rates and the latencies for the previously miscategorized and previously properly categorized ambiguous colour trials in the self- and the other-perspective blocks, separately for the geometric figures and the animal pictures. Analysis of the other-perspective blocks indicated a significant difference between the two conditions both for the geometric figures of the implicit transfer task (Z=-2.83, p= .005, r=0.485) and the animal pictures of the anticipatory looking task (Z=-3.07, p= .002, r=0.523). As can be seen from the figure, hit rates were markedly lower for the previously miscategorized than for the previously properly categorized ambiguous colour trials when participants had to categorize the items from their partner's perspective, despite the high accuracy on the same trials in the self-perspective blocks (miscategorized condition hit rate - figures: self > other, Z=-3.92, p< .001, r=0.672; pictures: self > other, Z=-4.22, p< .001, r=0.724). There was no difference between the two conditions on the self-perspective trials (figures: Z=-0.39, p= .694, r=0.067; pictures: Z=-0.85, p= .395, r=0.146). These results indicate that participants were aware of and, when instructed, could take into account the fact that their partner perceived the previously miscategorized colour in a different way than they did. The individual variation was, again, very high: whereas 11 participants categorized the geometric figures according to the other's (updated) perspective on all eight miscategorized trials (32.35%), 12 did not show any sign of perspective-taking (35.29%) with the rest (n=11) doing so on one to seven trials (32.35%). As in Experiment 1, Spearman' correlation indicated a strong relationship between the hit rate of the figure and picture trials (other-perspective blocks - previously miscategorized condition: r_s=.938, p<.001). Importantly, latency analyses revealed that it took significantly more time for participants to categorize the previously miscategorized items from their own perspective than those that were ambiguous but were previously properly categorized by their partner, implying an interference from the other's (updated) perspective (figures: Z=-2.49, p= .013, r=0.427; pictures: Z=-2.51, p= .012, r=0.431). For the other-perspective trials, the difference between the two conditions was not significant (figures: Z=-0.81, p= .417, r=0.138; pictures: Z=-0.47, p= .638, r=0.081).

With respect to the post-test questionnaire: 29 participants reported noticing the change in the other's behaviour and 25 provided a mentalistic explanation for their 'partner's' miscategorization at the end of the experiment.

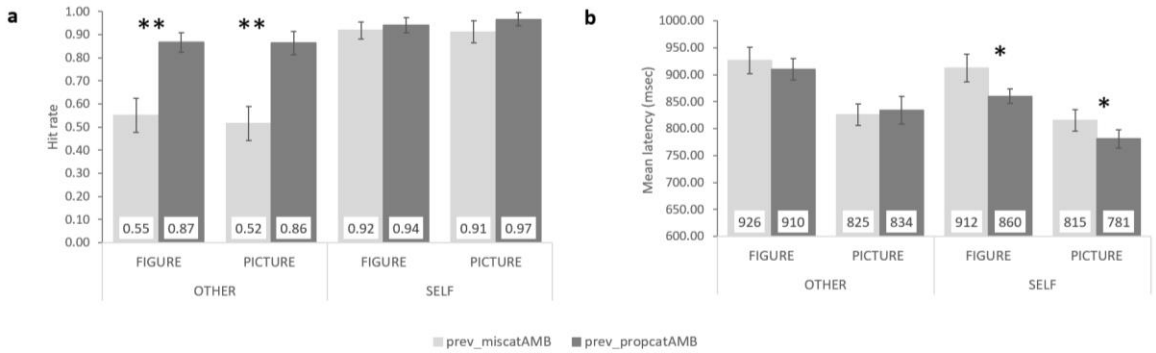


Figure 2.10. (a) Mean hit rate) and (b) latency on the previously miscategorized (prev_miscatAMB) and previously properly categorized ambiguous colour trials (prev_propcatAMB) of the explicit perspective-taking task, in Experiment 2. FIGURE denotes the geometric figures used in the implicit transfer task, PICTURE refers to the animal pictures used in the anticipatory looking task. OTHER vs SELF: blocks in which participants had to categorize the items from their 'partner's' versus their own perspective. Lower hit rate on the previously miscategorized trials in the OTHER blocks reflect perspective-taking. Latency is the time elapsed from the appearance of the cursor until the participant clicked on the figure. Error bars represent SE.
+: p<0.1; *: p<0.05; **: p<0.01.

2.3.2.5. Subgroup analyses based on the performance on the explicit task

Splitting the sample along participants' performance on the other perspective trials, based on the same criteria as in Experiment 1, yielded an n=19 'update' (7 males, age: M_{age}= 22.84, SD_{age}=4.17) and an n=15 'noupdate' participants (10 males, age: M_{age}= 25.07, SD_{age}=5.13). The two subgroups did not differ in age: $t(32)$=-1.39, p= .172, d=0.477) or gender (Fisher's Exact Test: p= .166).

2.3.2.5.1 Update subgroup

Anticipatory looking task: proportion of looking (test phase)

Figure 2.11a shows the mean proportion of looking for the miscategorized and the properly categorized ambiguous condition in the four blocks of the test phase for the 'update' subgroup of Experiment 2. Analysis of the proportion of looking data yielded a significant main effect of condition (Wald χ^2=43.98, df=1, p< .001), block (Wald χ^2=20.14, df=3, p< .001) as well as a significant condition x block interaction (Wald χ^2=26.82, df=3, p< .001). Whereas the proportion of looking towards the incorrect box on the properly categorized ambiguous trials did not change substantially over time, the proportion of looking towards the target box increased sharply from the first to the second block in

the miscategorized condition and continued to increase steadily throughout the rest of the trials. Pairwise comparisons revealed that difference between the two conditions was significant already from the first block (miscategorized>properly categorized: block1 – z=3.54, M_{diff}= 0.25, Wald 95% CI [0.11-0.38], p_{adj}<.001; block2 – z=5.26, M_{diff}= 0.44, Wald 95% CI [0.28-0.61], p_{adj}<.001; block3 – z=7.19, M_{diff}= 0.55, Wald 95% CI [0.40-0.70], p_{adj}<.001; block4 – z=6.87, M_{diff}= 0.59, Wald 95% CI [0.42-0.76], p_{adj}<.001), indicating that members of the 'update' subgroup quickly updated the other's mental state and revised their expectations regarding his/her behaviour after witnessing him/her acting in a way that was incompatible with their original assumptions. As for the first anticipatory period, participants generally looked more towards the target box in the miscategorized than towards the incorrect box in the properly categorized ambiguous condition from the beginning of the task (see Supplementary Materials **Figure S2.3b**).

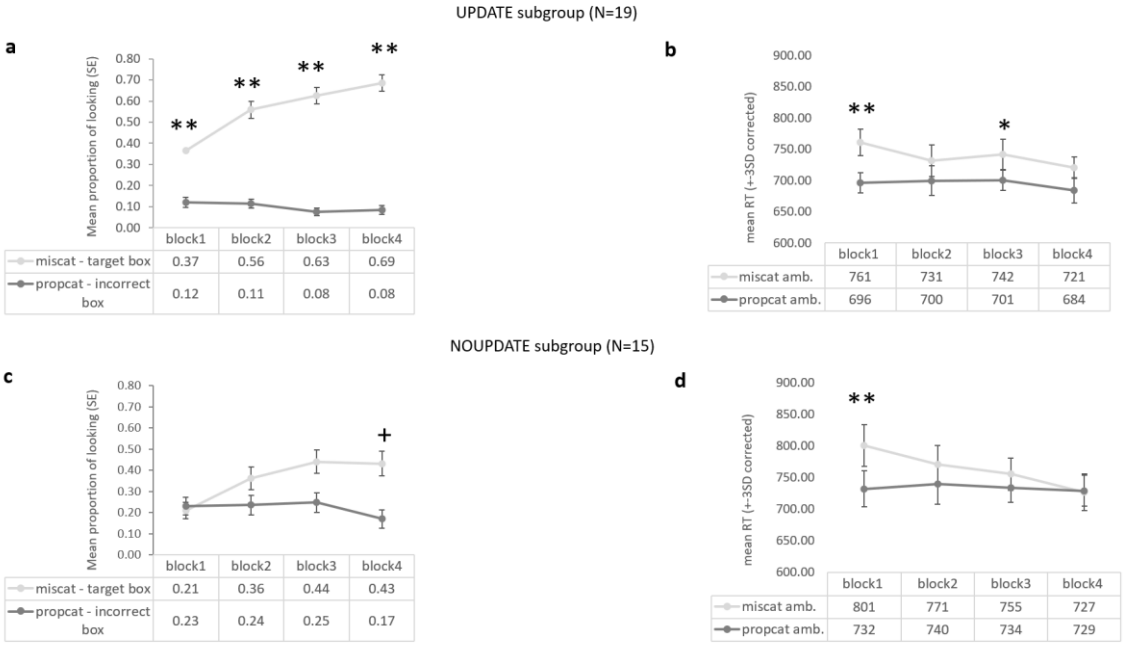


Figure 2.11. Changes in the proportion of looking towards the target box in the miscategorized condition (miscat: light grey line) versus the incorrect box in the properly categorized ambiguous condition (propcat: dark grey line), prior the partner's box selection, and (b) in mean reaction times in the miscategorized (light grey line) and properly categorized ambiguous (dark grey line) conditions in the (a, b) UPDATE (upper panel) and (c, d) NOUPDATE (lower panel) subgroups of Experiment 2. The subgroups were created on the basis of participants' performance on the other-perspective trials of explicit perspective-taking task. The figure displays the raw data. Error bars represent SE. +: p_{adj}<0.1; *: p_{adj}<0.05; **: p_{adj}<0.01.

Anticipatory looking task: reaction time

Analysis of the reaction times revealed a significant main effect of condition, $F(1,11)=22.53$, $p< .001$, $\eta_p^2= .556$, resulting from generally higher reaction times in the miscategorized condition, but only a tendency level main effect of block, $F(3,54)=2.25$, $p= .093$, $\eta_p^2= .111$), due to a slight decrease in reaction times in the last block, in both conditions, and no significant condition x block interaction, $F(3,5)=0.89$, $p= .453$, $\eta_p^2= .047$. Although participants were slower on the first few miscategorized trials than on the subsequent ones (see **Figure 2.11b**), indicating that they were initially somewhat surprised by the other's unexpected actions, pairwise comparisons revealed no significant difference between the first and the other blocks, after adjusting for multiple comparisons (all ts<2.26, and p_{adj}s > .110). There was no difference between the first and the subsequent blocks in the properly categorized ambiguous condition either (all ts<0.930, and ps > .364). As can be seen on the graphs, participants were much slower on the miscategorized than on the properly categorized ambiguous trials in the first block ($t(18) = 4.22$, $p_{adj} =.004$, d=0.970) and, to a lesser extent, also in the third block ($t(18) = 3.17$, $p_{adj}= .020$, d=0.727) but not in the second block ($t(18) = 1.49$, $p_{adj}= .616$, d=0.341). Interestingly, a tendency level difference remained between the two conditions even by the last block ($t(18) = 2.49$, $p_{adj}= .092$, d=0.571), suggesting that 'update' participants could not fully adapt to the change in the other's behaviour.

Implicit transfer task

With respect to the implicit transfer task, the analyses revealed a significantly lower hit rate for the previously miscategorized than for the previously properly categorized ambiguous colour trials (hit rate: Z=-2.45, $p= .014$, r=0.562). As can be seen from **Figure 2.9a** (middle panel) 'update' participants, on average, took into account the other's different perspective on roughly 50% of the trials. Indeed, most of the participants (68.43%%) took into account the partner's perspective to some degree, with only six participants (31.57%) not doing so at all.

Despite the numerically longer response latencies on the previously miscategorized colour trials (854 vs 828 ms), the difference between the two conditions in terms of latencies was not significant (Z=-1.45, $p= .147$, r=0.333).

Explicit perspective-taking task

Regarding the explicit perspective-taking task, although participants were slower in categorizing the miscategorized than the previously properly categorized figures from the partner's perspective, this difference was not significant (Z=-1.21, p= .227, r=0.285). No such difference was present for the animal pictures (Z=-0.77, p= .445, r=0.182). Analyses indicated significantly longer latencies for the previously miscategorized compared to the previously properly categorized items on the self-perspective trials (figures: Z=-2.21, p= .027, r=0.507; pictures: Z=-2.25, p= .024, r=0.516), implying an interference from the other's (updated) perspective. With respect to hit rate, the difference between the two conditions was not significant (figures: M_{diff}=0.04, Z=-0.59, p= .553, r=0.135; pictures: M_{diff}=0.10, Z=-0.76, p= .450, r=0.172, see **Figure 2.12a** and **1.12b**). At the end of the experiment, 17 out of the 19 participants could specify exactly how the other's behaviour deviated from the expected, and 16 gave mentalistic accounts for the observed change.

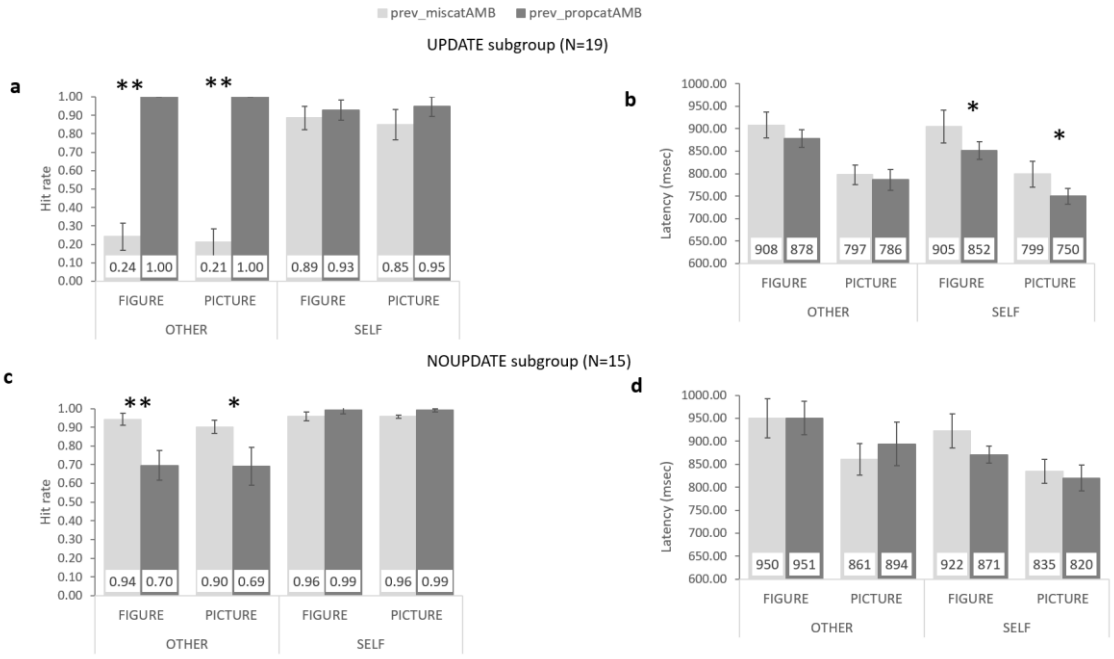


Figure 2.12. Mean hit rate (a, c) and latency (b, d) on the previously miscategorized (prev_miscatAMB) and previously properly categorized ambiguous trials (prev_propcatAMB) of the explicit perspective-taking task, in the 'update' (upper panel) and 'noupdate' (lower panel) subgroups of Experiment 2. FIGURE denotes the geometric figures used in the implicit transfer task, PICTURE refers to the animal pictures used in the anticipatory looking task. OTHER vs SELF: blocks in which participants had to categorize the items from their partner's versus their own perspective. Lower hit rate on the previously miscategorized trials in the OTHER blocks reflect perspective-taking. Error bars represent SE. Note that the figure also displays the hit rate on the OTHER trials, to provide a full picture of the data pattern. However, as we used the hit rate on the previously miscategorized and properly categorized trials to create the two subgroups, we do not compare these trials statistically.
+: p<0.1; *: p<0.05; **: p<0.01.

2.3.2.5.2 Noupdate subgroup

Anticipatory looking task: proportion of looking (test phase)

Figure 2.11c presents the mean proportion of looking in the four blocks of the test phase for the two conditions in the 'noupdate' subgroup of Experiment 2. Analysis of the proportion of looking data revealed a tendency level main effect of condition (Wald χ^2=3.36, df=1, p= .067) but no significant main effect of block (Wald χ^2=5.96, df=3, p= .113). There was, however, a significant condition x block interaction (Wald χ^2=14.82, df=3, p= .002). While the proportion of looking towards the target box in the miscategorized condition started to increase after the first few trials, the proportion of looking towards the incorrect box in the properly categorized ambiguous condition tended to decrease towards the end of the task, suggesting that, to some extent, 'noupdate' participants also revised their expectations regarding their partner's behaviour. Despite the marked difference between the two conditions in the third and the fourth block, pairwise comparisons indicated only a marginally significant difference between the two, and only in the fourth block, after adjusting for multiple comparisons (block3 – z=2.12, M_{diff}= 0.19, Wald 95% CI [0.01-0.37], p_{adj}= .136; block4 – z=2.33, M_{diff}= 0.27, Wald 95% CI [0.04-0.50], p_{adj}= .080). With respect to the first anticipatory looking period, although participants tended to look more towards the target box on the miscategorized than towards the incorrect box on the properly categorized ambiguous trials, this difference was not significant (for details see the Supplementary Materials).

Importantly, unlike in Experiment 1, the 'noupdate' subgroup, anticipated significantly less than the 'update' subgroup on both the miscategorized (M_{update}=0.39, SD_{update}=0.24 vs $M_{noupdate}$=0.47, $SD_{noupdate}$=0.26, Z=-2.84, p=.005, r=0.487) and the properly categorized ambiguous trials (M_{update}=0.13, SD_{update}=0.17 vs $M_{noupdate}$=0.45, $SD_{noupdate}$=0.32, Z=-2.88, p=.004, r=0.494), suggesting that they may have been less motivated or able to predict the other's actions, in general (for details see the Supplementary Materials, **Table S2.2**).

Anticipatory looking task: reaction time

Analysis of reaction time data yielded a significant main effect of condition, F(1,14)=10.59, p= .006, η_p^2= .431, but no significant main effect of block, F(1.90,26.55)=1.56, p= .230, η_p^2= .100. There was, however, a significant condition x block interaction, F(3,42)=3.85, p= .016, η_p^2= .216, resulting from the fact that, after the first few trials, 'noupdate' participants' reaction times decreased steadily in the miscategorized condition (see **Figure 2.11d**). Pairwise comparisons revealed a marginally significant

difference between the first and the last block ($t(14)= 2.68$, $p_{adj}= .054$, $d=0.675$; for all other comparisons with the first block: ts<1.83, and , p_{adj}s > .270) and the second versus the last block ($t(14)= 2.61$, $p_{adj}= .060$, $d=0.691$; all other ts<2.18, and p_{adj}s > .144), after adjusting for multiple comparisons. There was no such change in the reaction times over the course of trials in the properly categorized ambiguous condition (all ts<0.693, and ps > .499). Pairwise comparisons revealed that the difference between the miscategorized and properly categorized ambiguous condition was significant in the first block ($t(14)= 5.28$, $p_{adj}< .001$, $d=1.363$) but not in the subsequent blocks (all ts <2.01, p_{adj}s> .259), implying a relatively fast behavioural adjustment to the change in the other's behaviour.

Implicit transfer task

With respect to the implicit transfer task, just like 'update' participants 'noupdate' participants were also slower on the previously miscategorized than on the previously properly categorized ambiguous colour trials (see **Figure 2.9b**), with the difference being significant in this subgroup ($t(14)= -2.52$, $p= .041$, $d=0.581$), indicating an interference from the other's (different) perspective. Hit rates, on the other hand, were close to ceiling in both conditions, with no difference between the two ($Z=-1.21$, $p= .228$, $r=0.313$). In fact, there were only 2 participants in this group who demonstrated any sign of spontaneously taking the partner's perspective and acting accordingly (they made 1 or 2 'errors' i.e., passed shapes in line with the other's perspective, on the miscategorized but none on the properly categorized ambiguous trials).

Explicit perspective-taking task

Regarding the explicit perspective-taking task, 'noupdate' participants were equally fast in categorizing the previously miscategorized and properly categorized ambiguous figures (Z -0.23, $p= .827$, $r=0.061$) and they were slower when categorizing the previously properly categorized than the previously miscategorized animal pictures, from their partner's perspective, though this latter difference was not significant (Z -1.59, $p= .112$, $r=0.425$). With respect to the self-perspective blocks, participants' hit rate was at ceiling both for the previously miscategorized and properly categorized items, with no difference between the two conditions (figures: $M_{diff}=0.003$, $Z=-0.137$, $p= .891$, $r=0.035$; pictures: $M_{diff}=0.00$, $Z=-0.00$, $p=1.00$). Although they were slower on the previously miscategorized (figures: M=922 ms, SD=145 ms; pictures: M=835 ms, SD=103 ms) than on the previously properly categorized

ambiguous self-perspective trials (figures: *M*=871 ms, *SD*=73 ms; pictures: *M*=835 ms, *SD*=110 ms, see **Figure 2.12d**), the difference between the two conditions was not significant (figures: *Z*=-1.53, *p*= .125, *r*=0.395; pictures: *Z*=-0.37, *p*= .712, *r*=0.096). Despite the fact that 'noupdate' participants did (or could) not explicitly take the other's perspective, upon receiving the post-test questionnaire, 10 could specify how their partner's behaviour deviated from the expected, with 9 providing a mentalistic account for the partner's observed change.

2.3.2.6. Comparison of Experiment 1 and Experiment 2

2.3.2.6.1 Anticipatory looking task: proportion of looking

Familiarization phase

Comparison of the two experiments with respect to the proportion of looking in the familiarization phase indicated no difference between the two groups (later miscategorized ambiguous trials: M_{Exp1}=0.91 vs M_{Exp2}=0.85, *Z*=-0.50, *p*= .619, *r*=0.063; later properly categorized ambiguous trials: M_{Exp1}=0.82 vs M_{Exp2}=0.84, *Z*=-0.86, *p*= .388, *r*=0.108).

Test phase

To investigate whether the two experiments differed significantly in terms of how anticipatory looking changed in the two conditions over the test trials, in our main window of analysis, we ran an additional GEE for the proportion of looking data in the second anticipatory period, with condition, block and experiment as a predictor.

Comparison of the two experiments along the proportion of looking data in the test phase revealed a significant main effect of condition (Wald χ^2=55.48, df=1, *p*<.001), block (Wald χ^2=27.79, df=3, *p*<.001) and a significant condition x block interaction (Wald χ^2=69.18, df=3, *p*<.001), but no significant main effect of experiment (Wald χ^2=0.518, df=1, *p*= .472). There was, however, a tendency level experiment x condition interaction (Wald χ^2=2.84, df=1, *p*= .092), resulting from the fact that correct anticipation was generally higher on the miscategorized trials in Experiment 2 than in Experiment 1. Although the difference between the two conditions emerged earlier and was more pronounced by the last block in Experiment 2 than in Experiment 1, neither the experiment x block (Wald χ^2=1.60, df=3, *p*= .659) nor experiment x condition x block interaction was significant (Wald χ^2=2.80, df=3, *p*= .424), reflecting that,

the general pattern of the how proportion of looking evolved over the course of trials, was similar in the two experiments.

Performing the same analyses for the two 'update' subgroups yielded a significant main effect of condition (Wald χ^2=70.63, df=1, p<.001), block (Wald χ^2=3051, df=3, p<.001) and a significant condition x block interaction (Wald χ^2=56.67, df=3, p<.001), but no significant main effect of experiment (Wald χ^2=0.01, df=1, p= .940), experiment x condition (Wald χ^2=0.36, df=3, p=.551) or experiment x block interaction (Wald χ^2=1.05, df=3, p=.790). Importantly, however, there was a tendency level experiment x condition x block interaction (Wald χ^2=6.50, df=3, p= .090), resulting from the fact that the difference between the two conditions emerged earlier in Experiment 2 (first block) than in Experiment 1 (second block).

2.3.2.6.2 Anticipatory looking task: reaction times

To test whether the two experiments differed significantly with respect to how reaction times changed over the course of test trials (i.e. our measure of behavioural adaptation) we ran an additional 3-way ANOVA, with condition and block as within-subject and experiment as a between-subject factor. Comparison of the two experiments along the reaction times, yielded a significant main effect of condition, F(1,62)=65.74, p< .001, η_p^2= .515, block, F(3, 186)=6.51, p< .001, η_p^2= .095, and a significant condition x block interaction, F(3, 186)=11.20, p< .001, η_p^2= .431, but no significant main effect of experiment, F(1,62)=0.28, p= .596, η_p^2= .005, or experiment x block interaction, F(3, 186)=0.29, p= .993, η_p^2= .000. There was, however, a significant experiment x condition interaction, F(1, 48)=4.31, p= .02, η_p^2= .065, resulting from the fact that participants were much slower on the properly categorized ambiguous trials in Experiment 2 than in Experiment 1. Although visual inspection of the data indicated a somewhat larger drop in reaction times after the first few miscategorized trials in Experiment 1 than in Experiment 2 (Exp1: $M_{block1\text{-}block2}$=39.69 ms versus Exp2: $M_{block1\text{-}block2}$=29.67 ms), the experiment x condition x block interaction was not significant, F(3, 186)=0.97, p= .407, η_p^2= .015, indicating that the way reaction times changed in the two conditions over the course of trials was similar in Experiment 2 and Experiment 1.

Performing the same analyses for the two 'update' subgroups yielded similar results: a significant main effect of condition, F(1,29)=21.72, p< .001, η_p^2= .428, a tendency level main effect of block, F(3, 87)=2.27, p= .086, η_p^2= .073, and a significant condition x block interaction, F(3, 186)=7.48, p< .001, η_p^2= .205, but no significant main effect of experiment, F(1,29)=0.218, p= .644, η_p^2= .007, or experiment x block interaction, F(3, 87)=0.207, p= .891, η_p^2= .007. There was, however, a tendency level experiment x condition interaction, F(1, 29)=3.18, p= .085, η_p^2= .099, due to the fact that 'update'

participants were somewhat slower on the properly categorized and faster on the miscategorized ambiguous trials in Experiment 2 than in Experiment 1. Although this latter difference was present especially in the first block of the test phase, the experiment x condition x block interaction was not significant, $F(3, 87)=0.23$, $p= .875$, $\eta_p^2= .008$**.**

2.3.2.6.3 Implicit transfer task

To investigate whether our manipulation had a significant effect on participants' performance on the implicit transfer task, we computed the difference of the previously properly categorized ambiguous and miscategorized trials' hit rate (as well as latency) for each participant and ran pairwise comparisons for the difference scores of the two experiments. Despite the somewhat lower hit rates in Experiment 2, compared to Experiment 1, on the previously miscategorized colour trials, the two groups did not differ significantly with respect to the magnitude of difference between the two conditions, i.e. in the extent to which participants took into account the other's different perspective while performing the task (Z=-0.12, p= .905, r=0.015). There was, however, a significant difference between the two experiments with respect to latencies (Welch's $t(52.06)= -2.48$, p= .017, d=0.494), resulting from the fact that, unlike in Experiment 1, in Experiment 2 it took significantly longer for participants to make their decisions on the previously miscategorized compared to the previously properly categorized ambiguous colour trials. This suggests that even though they rarely took the other agent's (different) perspective, participants in this experiment did consider it, when passing the figures.
When comparing the two 'update' subgroups only, the differences were not significant (hit rate difference: Z=-0.72, p= .471, r=0.129; latency difference: Z=-1.42, p= .156, r=0.255).

2.3.2.6.4 Explicit perspective-taking task

Performance on the explicit perspective-taking task was compared by computing difference scores in a similar way as above (previously properly categorized ambiguous - miscategorized trials hit rate or latency), separately for the other- and self-perspective blocks, to obtain a measure for the extent of explicit perspective-taking and interference from the 'partner's' perspective, respectively.
With respect to the self- perspective blocks, participants in Experiment 2 were marginally slower (Z=-1.92, p= .054, r=0.244) in categorizing the previously miscategorized colour (compared to the previously properly categorized ambiguous colour) geometric figures than in Experiment 1, suggesting

a stronger interference from the other's (updated) perspective in this experiment. A similar pattern was observed for the animal pictures with the previously miscategorized colour: participants in Experiment 2 were significantly slower in making the decision for the previously miscategorized colour (compared to the previously properly categorized ambiguous colour) animal pictures (Exp1 M_{diff}=18.43 ms vs Exp2 M_{diff}=-33.93 ms, Z=-3.24, p= .001, r=0.412), compared to participants in Experiment 1, indicating a larger impact of the other's perspective on their own in Experiment 2 than in Experiment 1. The two experiments did not differ in terms of how many errors participants made when categorizing the previously miscategorized (compared to the previously properly categorized) items (geometric figures: Z=-.68, p= .500, r=0.086; animal pictures: Z=-0.38, p= .706, r=0.048). Similar findings were obtained when comparing only the two 'update' subgroups, along the difference scores computed for the self-perspective blocks. Despite 'update' participants being slower in Experiment 2 than in Experiment 1 both on the previously miscategorized (compared to the previously properly categorized ambiguous colour) geometric figures and animal pictures the difference between the two experiments was, significant only for the animal pictures (geometric figures: Z=-1.42, p= .155, r=0.255; animal pictures: Z=-2.37, p= .018, r=0.426).Regarding the other-perspective blocks there was no significant difference between the two experiments in the extent of perspective-taking (hit rates – figures: Z=-1.13, p= .259, r=0.144; pictures: Z=-1.31, p= .191, r=0.166; latencies - figures: Z=-0.76, p= .445, r=0.096; pictures: Z=-0.24, p= .810, r=0.003), despite the difference between the two conditions being generally larger in Experiment 2 than in Experiment 1. The same was the case when comparing the two 'update' subgroups only (hit rates – figures: Z=-0.25, p= .807, r=0.044; pictures: Z=-0.15, p= .884, r=0.026; latencies - figures: Z=-0.12, p= .906, r=0.021; pictures: Z=-0.08, p= .937, r=0.014).

2.3.3 Discussion

Experiment 2 replicated and extended the findings of Experiment 1, by showing that when two potential sources of interference were eliminated (the explicit categorization rule and the presence of colour labels, evoking strong colour label associations), correct anticipations tended to emerge earlier on the miscategorized trials, and seemed to be more pronounced throughout the experiment, specifically in the 'update' subgroup. The ratio of participants, for whom there was explicit evidence for updating the other's mental state, was also slightly higher in this than in the previous experiment. Interestingly, the higher correct anticipations were not accompanied by a more marked behavioural adjustment. In fact, the pace at which reaction times decreased after the first block on the

miscategorized trials was even somewhat slower, and the difference between the two conditions was generally less pronounced in Experiment 2 than in Experiment 1.

Importantly, the extent of explicit perspective-taking, was roughly the same in the two experiments. There was no significant difference between the two experiments in the extent of spontaneous perspective-taking on the implicit transfer task either. This does not mean, however, that the degree to which participants considered their partner's (updated) perspective following the anticipatory looking task was the same in the two experiments. Unlike in Experiment 1, in Experiment 2, the other's perspective seemed to interfere with participants' responses, both in the implicit transfer task and when they had to explicitly categorize the previously miscategorized colour from their own perspective as indicated by the longer latencies in both cases. Interestingly, such an interference from the other's perspective was present not only in the 'update' but also in the 'noupdate' subgroup. This suggests that the 'noupdate' participants might have been aware of the difference in the two perspectives to some extent, but they just did not deploy the information they represented, either because it was not consciously accessible or because they were uncertain whether they drew the correct conclusion regarding the other.

2.4 General Discussion

The aim of the present study was to investigate whether people spontaneously update the content of another agent's mental state and revise their expectations about her future actions, upon observing a behaviour that does not correspond to their original assumptions regarding the beliefs the other may hold, even when this updating would not be necessary in the given situation. To this end, we applied a virtual referential communication task and eyetracking methodology. In specific, in two experiments we investigated how anticipatory looking develops when a 'partner', who performs a colour categorization task, starts to categorize an ambiguous colour in an unusual way, different from how she did before, making updating her mental state necessary, to be able to explain and predict her behaviour. In Experiment 1 the 'partner's' miscategorization was a violation of an explicit categorization rule (e.g. 'put blue objects into the box labelled as blue'). In Experiment 2 the miscategorization meant a change in how the other judged the similarity of two colours (to which box the picture belongs to based on the frame's colour). We assumed that taking and updating the other's perspective will be easier in Experiment 2 as a similarity judgement may prompt participants to

consider the partner's subjective evaluation more than compliance with an explicit rule. To ensure that we measure a spontaneous updating of the other's mental states, the task participants performed did not require predicting their 'partner's' actions, as they had to simply click on the box that lit after the partner made a choice. Besides investigating how people revise their expectations about other's perspective, we also aimed to assess whether they use the information about the other agent's updated mental state spontaneously in subsequent social interactions. To this end, following the anticipatory looking task we also administered virtual coordination task to our participants, in which participants could implicitly transfer their newly acquired knowledge about how the other sees a particular colour in order to be more efficient in achieving a shared goal.

To summarize the main findings of the anticipatory looking task, in both experiments we found that participants revised their expectations regarding the other's behaviour spontaneously, without any external prompt. Correct anticipations emerged and started to differ significantly from what could be expected by random looking soon after the first few miscategorized trials, both when updating required participants to overcome a supposedly strong interference from an explicitly set categorization rule and an established colour-colour label association (Experiment 1) and also when such explicit categorization rule was not present (Experiment 2). The effect manifested not only in how much participants looked towards the correct location but also in terms of where they directed their initial looks.

Importantly, for approximately half of the participants, in both experiments, performance on a subsequent explicit perspective-taking task provided evidence that such a change in the anticipatory looking behaviour reflected the updating of the other's mental state and not merely an overwriting of previously learnt stimulus-outcome associations. These participants could categorize items with the previously miscategorized colour from their partner's point of view, taking into account that the other's perspective differs from their own, even though they never saw the partner acting on those items, when they were asked to do so. In fact, the effects that emerged at group level were driven mostly, though not exclusively, by this subgroup. It is worth noting, that within these mental state update subgroups, by the end of the task (in Experiment 1 and even earlier in Experiment2), correct anticipations started to emerge not only prior the partner's action (anticipatory period 2 used for the main analysis) but also after the onset of the target word (anticipatory period 1, Supplementary Materials), corroborating previous findings which indicated an early sensitivity to the other's perspective in visual world paradigms (see e.g. Cane et al., 2017). In specific, these results suggest that if the output of the inferential process is consciously accessible and/or people are confident that their conclusion regarding the other's mental state is correct, human adults might make use of the information about the other's mental state to predict her behaviour from the earliest moment it is possible.

Although the general pattern of results was similar in the two experiments, in all three tasks, there were two differences that are important to point out. First, in the anticipatory looking task of Experiment 2, in which the inhibitory demands of the task were presumably lower (due to the lack of an explicit initial rule - "*green goes to the green box*" - that had to be ignored on the update trials), signatures of updating emerged earlier than in Experiment 1 (in block 1 versus block 2). This indicates that, just like other implicit ToM processes (see e.g. Schneider, Lam, et al., 2012), the spontaneous updating of other agents' mental states might also be dependent on the availability of executive resources. In line with this, more participants gave a mentalistic account of their partner's unusual behaviour in their verbal reports in the debriefing phase of this second than of the first experiment. The lack of initial explicit rule also had another effect: a somewhat smaller 'surprise' upon seeing the first few miscategorized trials in Experiment 2 than in Experiment 1 and a generally smaller difference between the two ambiguous conditions in terms of reaction times, from the beginning of the anticipatory looking task, due to responding more slowly also on the properly categorized ambiguous trials. This was most likely the result of participants' weaker initial expectations (and the resulting higher level of uncertainty) regarding how their 'partner should act. Second, although members of the 'update' subgroup did not confuse the other's perspective with their own representation of the colour, as evidenced by the low number of errors on the self-perspective trials of the explicit perspective-taking task, response latency results suggest that their own first-person judgements were nevertheless affected by their 'partner's perception of the previously miscategorized colour. Humans' propensity to take into account others' visual perspective spontaneously (called 'altercentric intrusion effect') has been widely demonstrated in studies, where another agent is present who holds a discrepant perspective (Samson et al., 2010; Surtees, Apperly & Samson, 2016). Latency results from Experiment 2 suggest that such an altercentric intrusion effects may even occur in the physical absence of another agent and may be triggered by the inference about how the invisible other may perceive certain elements of the environment.

Members of the 'update' subgroups also took into account what they had learnt about their 'partner' (during the anticipatory looking task) in the implicit transfer task, at least to some extent, suggesting that if human adults realize that another agent's perspective differs from their own, in some respect, they make use this information spontaneously later, when interacting with the other. Interestingly, however, they acted in accordance with the other's perspective on less than 50% of the trials, passing the geometric shapes with the miscategorized colour at the gates which corresponded to their own (and not to their 'partner's') perception, in most of the cases. They acted so despite knowing that these items had the same colour as the pictures frames that were miscategorized earlier by the other, as indicated by the strong correlation between their performance on the figure and the picture trials of the explicit perspective-taking task later. The implicit transfer of the updated knowledge from one task

to the other remained strikingly low even after eliminating two possible sources of interference from Experiment 1, the presence of colour labels and the explicit categorization rule. While these effects might have stemmed from a strong egocentric-bias and/or human adults' low propensity to generalize newly acquired knowledge about the others' belief contents to subsequent social interactions, it might have also been the consequence of certain, specific features of the task. Note that in this task participants had to pass novel geometric shapes which had the same colours as the picture frames in the anticipatory task through blue and green gates to the partner who had to put the item in a box in the next step, which they did not see. To maximize joint efficiency (speed) we assumed that it would be beneficial to consider the other's perspective and pass the object at the gate where she might expect it. Since participants did not see their 'partner's' actions and were not given feedback about her success/failure either, they might have thought that joint efficiency could be achieved without considering the other's perspective, either because they thought the partner is fast enough (so that it does not matter at which gate they pass the figure) or because they expected the other to adjust to them. After all, successful coordination requires mutual adaptation of the partners and co-efficiency rarely depends on only one member of a pair. In addition, behavioural adjustment might also require the people to be highly certain in that their attribution is correct, which might not have been the case in our experiments, given that participants never received any explicit feedback that the interpretation they had come up with in the anticipatory looking task is actually correct.

In addition to our main results, the early emerging correct anticipation; an ability to explicitly take the other's updated perspective and take it into account spontaneously - to some extent - our findings also reveal a strong egocentric bias, in both experiments, even in the 'update' subgroups. Such a bias was clearly present both in the anticipatory looking task, where the level of correct anticipation on the miscategorized trials remained significantly below the level observed on the other ambiguous trials, even after eliminating the rule-bias, and also in the explicit perspective-taking task, where even 'update' participants tended to make egocentric errors (roughly 20%) when judging the perspective of the other. It might have played an important role also in the low level of spontaneous transfer of the updated content, besides all the above-mentioned factors. Such results are generally in line with the findings showing that humans often fail to deploy their ToM abilities in online interactions and fail to take into account the other's differing visual perspective when performing a task (Keysar et al., 2000). Importantly, for roughly half of the participants there was no clear evidence for the updating of the 'partner's' perspective. These participants could not categorize the previously miscategorized colour items from the other's perspective when explicitly instructed to do so and/or made a large number of errors also when categorizing the other, previously properly categorized ambiguous colour, suggesting that they may have thought that the other is simply uncertain about the two ambiguous colours. It is an intriguing question how to explain the performance of this subgroup. One possibility is that

members of the 'noupdate' subgroups did not track the other's mental state at all during the anticipatory looking task. After all, in our paradigm it was possible to perform the task without doing so and these participants might not have been motivated to compute the other's beliefs, hence, might have chosen to follow a simple strategy to focus on only the lighting up of the box ('their task') and did not even try to predict the other's behaviour, or applied a nonmentalistic rule to predict the events (such as the 'greenish goes to blue box'). The fact that, in Experiment 2, the number of 'noanticipation' trials was significantly higher in these than in the 'update' subgroup supports the first explanation. If the interpretation is correct and members of the 'noupdate' subgroup indeed did not engage in mentalization, their results actually provide evidence for the non-automatic nature of spontaneous ToM: the computation (and consequently the recomputation) of other agents' belief content may not be triggered under all circumstances, in everyone (see Schneider et al., 2017).

Alternatively, 'noupdate' participants might have actually started to search for a mentalistic explanation upon observing their 'partner's' unexpected actions, but just did not manage to arrive at a satisfying or strong enough conclusion within the available time window or arrived to it by the very end of the task. After all, action interpretation is an ill-posed problem. An agent can have multiple reasons to act in the way she does and if an observer does not have access to sufficient information about the factors determining the other's behaviour at the moment when the unexpected action occurs, it might be difficult to narrow down the potential explanations. Importantly, in our task, there were no contextual cues to which an explanation could have been anchored, i.e. that could have helped interpreting why the other started to miscategorize the colour. In addition, miscategorized and properly categorized ambiguous colour trials were intermixed which could have made noticing of the behaviour change for a specific colour but not for others in itself hard. 'Noupdate' participants might have just thought for a longer time that the other is making random errors. The fact that correct anticipation did emerge in the 'noupdate' subgroups as well, just much later than in the 'update' subgroups, and that many of these participants could report how the other's behaviour deviated from the expected and provided a mentalistic account for the other's miscategorization at the end of the experiment, lends support to this second explanation, that 'noupdate' participants simply struggled more to find a proper explanation for the partner's observed behaviour. This may be especially true for the 'noupdate' participants of Experiment 2, who, besides starting to anticipate correctly by the end of the task, demonstrated clear signs of interference from the other's updated perspective on the previously miscategorized colour trials later on, in both the implicit transfer and the explicit perspective-taking task. 'Noupdate' participants were thus rather likely 'less successful updaters', but not in fact no updaters potentially because of the specific characteristics of our paradigm.

In conclusion, results from Experiment 1 and Experiment 2 provide the first evidence for the spontaneous updating of other agents' mental states on the basis of their observed behaviour. At the

same time, individual differences in how anticipatory looking developed on the miscategorized trials show that the process is by no means fully automatic and, depending on several factors, it might take shorter or longer time to update the content of the belief attributed to others. Future studies should investigate these factors in depth, extending the research to non-perceptual mental states, e.g. more general knowledge, interactive contexts and situations that are richer in contextual cues that may help participants to interpret the other's behaviour. Finally, future studies should also address the circumstances under which people adjust to the updated mental state of the other.

Chapter 3: Representation of other agents' hypothesis space

In the past 40 years, the majority of studies investigating humans' ability to reason about other agents' mental states, have focused on the capacity to understand that an agent might have an incorrect belief about reality, as a result of not witnessing certain events, and tested whether people can accurately predict that the agent will act on the basis of the false belief he holds. Studies addressed whether people compute the content of such false beliefs spontaneously (Schneider et al., 2017; Scott & Baillargeon, 2017), whether they can attribute false beliefs to other agents only about the location or also about the identity of objects (Low & Watts, 2013; Kampis & Kovács, 2022) and, more recently, whether they readily update the content of false beliefs if they learn that their original assumption was wrong or no longer hold - a question we investigated in the previous chapter.

Crucially, however, mentalizing entails much more than reasoning about false beliefs in change-of-location type of scenarios. Not only because the contents of other agents' beliefs are often much more complex than those that have been investigated extensively in the past, but also because agents, who lack a certain piece of information, for example, because they have not witnessed an event, are often not simply mistaken about the current state of affairs. They are not fully ignorant either. They are usually well aware of or at least suspect the fact that they have only partial knowledge about the actual state of the world, hence, what they may represent is seldom a single possibility with a high level of certainty, but rather a limited number of mutually exclusive alternatives, with some probability assigned to each. Given that we face situations of uncertainty on daily bases, such situations actually seem far more common in everyday life than false belief scenarios that have been the focus of ToM research for decades. For instance, a guard who lost sight of a thief they are chasing indoors, may not rely on a single guess about the next location to be checked, instead, he may represent multiple hypotheses regarding where the thief may be at the given moment, based on e.g. where one can exit the building ('*he either runs towards door A, B or C'*) and act accordingly, asking his co-workers to close all exits known by the public. Likewise, a taxi driver not knowing certain parts of the city by heart may have several ideas how a given destination should be approached, none of which may be completely wrong, though some may be better than others in terms of the length of the route ('*I shall either turn left here or continue my way straight ahead'*). Importantly, as observers or participants in these interactions we are also able to understand and foresee the possible actions these protagonists based on the possibilities they represent, just as the thief who expects the guards to first close all possible doors known by the public (but not the hidden maintenance doors he happens to know about) or the passenger who is aware of the driver's uncertainty when they arrive to a junction.

The ability to represent not only the actual state of affairs (*what is the case*), but also alternative states of the world (*what may be the case*), from first-person perspective, more specifically, what is possible under the laws of nature or probable to some extent, according to the person's current knowledge of the state of affairs (referred to as physical and epistemic possibility, respectively), at a certain

timepoint, plays a fundamental role in daily life. This allows humans to plan ahead, taking into account the present or future situational constraints, and select the most optimal course of action out of the various option available in the given context. In short, this enables preparation for the future (Redshaw, 2014; Redshaw & Suddendorf, 2016). Tracking what events other agents might consider possible and impossible, in situations, where they lack full knowledge about certain aspects of the world (for example where *exactly* someone or something is), can arguably yield similar benefits, via restricting the range of actions one can expect from the other agent in the near future. Encoding such information, i.e. the content of the other agent's hypothesis space, spontaneously, can extend the scope of contexts in which the other's actions become, to some extent, predictable. Thus, it may expand the range of situations in which humans can flexibly adapt to and smoothly interact with others. For instance, representing the alternatives the guards likely consider may help the thief to avoid running into their arms. Doing the same in case of the taxi driver, in the example above, allows the passenger to intervene in time before the driver would make a decision at a crossroad, that results in a longer route.

Although the representation of other agents' hypothesis space is undeniably beneficial to the observers, it is important to note that such beliefs differ from the ones entertained by the protagonists of the classic false belief task (over and above the apparent difference in the complexity of the content), in two crucial aspects. First, in some cases, it may be more difficult to assess the truth value of these belief contents than the truth value of belief contents represented by the other agent in standard ToM tasks. In situations where one of the alternatives represented by the agent corresponds to the actual state of affairs (for example the agent thinks that '*the object is either at location A or B*', and the object is actually at location A) the resulting belief is true according to the rules of logic, yet, contrary to classic 'true beliefs', it does not necessarily lead to a reality-congruent action[8] (since actions are often sequential, the agent might start to search at location B before turning to A). Second, independent of whether any of the alternatives are true (i.e. match the actual state of affairs), unlike the contents attributed in false belief tasks, which clearly determine the action one can expect from the agent, these types of belief contents do not render the agent's behaviour as predictable as simple true and false beliefs do (if the agent believes the object is either in box A or box B and assigns equal probability to the two alternatives, it is unpredictable where she will search first). To emphasize this

[8] Although in the example we provided the alternatives are exhaustive, such situations may also arise when the agent represents only some of the possible alternatives, if one of the alternatives the other represents happens to correspond to the actual state of the world,

feature of these beliefs, I will use the term 'underspecified' to refer to them in the following parts of the chapter[9].

Human adults can entertain mental state contents of virtually any type and complexity, in an explicit manner. Based on recent findings which suggest that implicit ToM processes rely on the same underlying mechanisms as explicit ToM (El Kaddouri et al., 2020; Hyde et al., 2015; Naughtin et al., 2017; Nijhof et al., 2016), one can assume that the spontaneous tracking of mental states also extends to belief contents that are not fully 'specified' (the way simple true and false beliefs are). A recent study by Kovács and colleagues (2021) with infants and another one by Hegedűs and Király (2022) with adults, seem to support this idea. In Kovács and colleague's (2021) study, 15-month infants first saw an agent hiding an object to an unspecified location (left or right) and leaving the scene. Afterwards, in the agent's absence, the location of the object was revealed to infants, and then it was again invisibly hidden, either in the agent's absence or presence (control condition). Given the second hiding, infants did not know where the object is. At this moment infants were allowed to search for the object. The authors hypothesized that if infants do not track the agent's beliefs they should search randomly, if, however, they encoded the agent's unspecified belief at the beginning (object at location x), the content of which they filled in later in the agent's absence (object in the left box) and they sustained this belief for the case the agent comes back, it might influence their search behaviour. Indeed, infants tended to search for the object at the location in which the agent believed the object to be in this (but not in a control) condition. This suggests that infants attributed an unspecified[10] belief content to the other agent about an object's location after the first hiding event, updated this content when they were provided with sufficient information about the object's location later and sustained a represented content even though they did not know its external validity, as if they had thought that it might be useful later. Using an object detection paradigm, which manipulated whether the agent had perceptual access to the potential location of an object (hidden from the observer), Hegedűs and Király (2022) found that such an unspecified belief content influences adult's behaviour in the same way it

[9] The representation of such belief contents is inherently accompanied by some degree of subjective uncertainty (as the observer cannot predict what the agent will do next). Most likely it also entails encoding of the agent's uncertainty in some sense. Nonetheless, labelling such a belief content 'uncertain' would be rather misleading, as it would imply an uncertainty regarding what alternatives the observed agent represents, which is definitely not the case if the constraints of the situation and the agent's knowledge clearly circumscribes hypohtesis space itself.

[10] The authors use the term underspecified in their paper but refer to a type of content Kovács (2016) labels as 'unspecified' (she believes that 'there is [something] [somewhere]'). Unspecified beliefs, unlike underspecified beliefs, do not grasp the alternatives potentially generated by the observed protagonist. To avoid confusion, I use the term 'unspecified' when describing the author's study here. Importantly, in Kovács and colleagues' study the observed protagonist had a clear representation about the object's location, the observer (the infant) just did not know exactly what that is. That it, it was the observer who was uncertain and not the observed person.

has been demonstrated to affect their responses for specified contents (see: Kovács et al., 2010), facilitating how fast participants detect the object.

Representing the actual content of another agent's hypothesis space (he thinks that *'the object is either at location A or location B'*), is likely to be cognitively more demanding than representing an unspecified belief content (such as '*the object is somewhere'*). It is also likely to be more demanding than representing the fact that the other '*does not know* the location of the object'. The reason is that, unlike unspecified beliefs or the representation of ignorance, these belief contents are logically structured propositions (propositions made up of multiple elements, possibly with logical operators between those). Nevertheless, when the agent's hypothesis space can be clearly circumscribed (i.e. it is obvious which options the agent may consider 'possible') and the number of hypotheses the other likely entertains is relatively low, the benefit of spontaneously representing what alternatives he/she considers in a particular situation, likely overrides the costs. It offers the opportunity to quickly intervene, if necessary, smoothly cooperate or efficiently compete with the other, by ensuring a 'preparedness' for the possible actions he/she may perform in the given context. In theory, all what this requires, besides the general capacity to readily attribute mental states to others, is the ability to co-represent multiple, mutually exclusive alternatives, spontaneously.

Given that the simultaneous representation of two or more incompatible possibilities necessarily involves setting up a disjunctive relation between those, according to most of the authors, research investigating humans' capacity to co-represent multiple alternatives has primarily focused on the question whether and how the target group being tested performs the logical inference called disjunctive syllogism or 'reasoning by exclusion' (*'if A OR B and NOT A, THEN (necessarily) B'*). Evidence suggests that adults represent disjunctions and reason by exclusion without having any intention or being aware of doing so. They can easily judge whether or not the final sentence of a story makes sense after reading scenarios the interpretation of which would not be possible without first representing a disjunction and then applying negation, tend to falsely believe that the conclusion they could draw while reading the story was explicitly presented as part of the text (Lea et al., 1990) and can quickly make lexical decisions about words semantically related to the proposition they could infer via disjunctive reasoning (Lea et al., 1995). Adults also demonstrate a pupil dilation pattern indicative of performing such computations while viewing scenarios in which ambiguity regarding the identity of an object can be resolved by applying disjunctive syllogism at a certain timepoint (Cesana-Arlotti et al., 2018). Recent findings, from looking time and pupillometry studies, suggest that this capacity might already be part of the preverbal infant's cognitive repertoire (Cesana-Arlotti et al., 2018; Cesana-Arlotti et al., 2020). In a series of experiments, using the same type of paradigm that was used with adults (in which disambiguation of a hidden object's identity became possible at some point on the basis of the presented information, by applying disjunctive syllogism) Cesana-Arlotti and colleagues (2018) found

that even 12-month old infants looked longer when the outcome presented at the end (e.g. snake exiting the occluder) was inconsistent with the logical inference they could draw before ('the hidden object must be the ball'), with their pupils dilating more when the scene licensed such an inference (as opposed to when it did not). Such a finding implies that infants represented both alternatives at the beginning, in the form of a disjunction (*'the object is either a snake or a ball'*), and performed the appropriate inference, spontaneously, when the evidence presented allowed them to eliminate one of the options[11]. The presence of these inferences, well before they would be able to produce or even understand basic logical words, like OR or NOT, or would master language in general, provides further evidence that humans may represent alternatives spontaneously, relying on language-independent, possibly innate cognitive processes.

Some findings seem to indicate that, just like the first-person representation of alternatives, the representation of other agents' hypothesis space may also take place spontaneously and rely on a mechanism that is present from very early on. In particular, the results of a control experiment of Knudsen and Liszkowski (2012b), which originally aimed to disentangle false-belief and ignorance-based accounts of infants' helpful communicative behaviour, may be interpreted as evidence for the presence of such a capacity, already in infants. In this control experiment, 18-month-old infants saw the hiding of a toy in one of two boxes, then, in the absence of the toy's owner, the toy was removed from the scene and boxes were baited with unpleasant materials (while the agent was still absent). In this case, participants pointed equally often at both boxes upon to agent's return, to warn her about the aversive materials and thereby help her to avoid getting in contact with them. This suggests that they understood that someone who has not witnessed where an object was hidden, out of two possible locations, will represent two alternatives, and hence will be equally likely to search for the object at both locations. Importantly, however, such a behaviour is also compatible with a more parsimonious account, that infants simply represented the other agent's ignorance regarding the toy's location ('she does not know [where it was hidden]'), consequently, did not form any specific expectations regarding where the agent will search for the desired toy. They pointed randomly at both boxes because they knew that the agent wants to find the toy and predicted that she will search at the available locations. Whether humans indeed represent alternatives spontaneously from third-person perspective, is therefore still an open question.

The present study aims to address this question directly, i.e. investigate whether human adults represent the content of other agents' hypothesis space, spontaneously, by testing whether another

[11] Although some authors question the interpretation of these findings, claiming that infants simply used 'serial guessing' (Leahy & Carey, 2020), instead of representing the two alternatives simultaneously, the pattern of results in a more recent study of the authors, using the same paradigm (Cesana-Arlotti et al., 2020), makes this highly unlikely. Importantly, as the authors point it out, even is infants would use such a guessing technique, it unclear how it could work without first representing the space of alternatives.

agent's underspecified belief content, regarding an object's location, influences the observers' spatial attention, in situations where it is not necessary to track the beliefs the other holds. This study essentially builds on two lines of research. First, a vast amount of data shows that peoples' own representation of a scene drives their attention to the corresponding elements of the external world (specifically, to their locations) even if they do not act on those (see e.g. Allopenna et al., 1998; Altmann & Kamide, 1999; Tanenhaus et al., 1995). In particular, several studies have found that when listening to a sentence, adults direct their attention to the object corresponding to the upcoming word (e.g. the cake and not other objects upon hearing the phrase 'the boy will eat'; Altmann & Kamide, 1999). There is evidence that such anticipatory eye movements do not simply reflect associations between the object and the word (Kamide et al., 2003), with a recent study showing that people's attention is indeed driven by what the sentence entails i.e. by their 'mental representation' of the actual state of affairs (as indicated by results showing that upon hearing a sentence about drinking in past tense people tend to look at the empty rather than the full cup that is being present; Altmann & Kamide, 2007). Second, numerous studies have demonstrated that, when another agent is present, people's actions, in particular their perceptual judgements, are affected by the visual perspective and the content of the (false) belief the other agent holds, even if the other is passive and his perspective is completely irrelevant for the task they perform (see e.g. Samson et al., 2010; and Kampis & Southgate, 2020 for a review). For instance, in a seminal study of adult implicit ToM, Kovács and colleagues (2010) found that the false belief of a task-irrelevant agent about an object's presence facilitates adults' object detection performance. Other studies, showing, for example, that participants who had to judge the visibility of low-contrast Gabor patches, were more likely to detect near-threshold stimuli when an avatar could also see them, indicate that the content of another agent's mental state may also influence adults' perceptual sensitivity (Seow & Fleming, 2019), though it is not clear whether this effect is specific to situations where the observer and the agent share visual perspective or also extends to those where the other holds a discrepant belief about the state of affairs.

Importantly, the content of another agent's (false) belief and visual perspective has been found to influence not only the decisions people make but also where their attention is directed, as indicated most clearly by the eye movement patterns they produce while watching false-belief scenarios in nonverbal ToM tasks (Schneider, Bayliss, et al., 2012; Schneider et al., 2014) or upon listening to a speaker's instructions, whose perspective differs from their own, in referential communication tasks (see e.g. Cane et al., 2017). Although recent studies failed to replicate some of these findings, specifically those obtained with implicit false belief tasks measuring anticipatory eye movements prior the protagonist's action (see: Kulke, von Duhn et al., 2018), altogether, the available evidence suggests that people's spatial attention may be influenced by not only their own representation of the actual state of the world but also other agents' representation of the current state of affairs.

Based on these considerations, we conjectured that in situations where the other agent is uncertain about the location of an object, people who have encoded the content of the agent's hypothesis space, should attend not only to those locations they themselves but also to those that only the other agent considers to be a 'possible' hiding place for the object. They should do this even if they know where the object has been hidden, i.e. the actual state of affairs, and they are certain in their knowledge. While it is unclear whether such an effect could be captured by measuring participants' anticipatory eye movements, given the apparent unreliability of this measure even if the other's behaviour is clearly predictable (Schuwerk et al., 2021), it might be present in other behavioural indices of spatial attention such as the speed and accuracy of the perceptual judgements people make about events occurring at the 'possible' location(s). To test whether this is the case, we developed a novel ToM task, in which an agent could either have a true or an 'underspecified' belief about the location of an object. There were four locations, differing in size and shape. We measured how participants detect changes at the four locations when the other agent represented multiple, equally likely, mutually exclusive alternatives, one of which always matched reality (and the participant's own knowledge). Such a design made it possible for us to validate our paradigm: test, whether performance on the task is influenced by – at least – the content of participants' own representation. Crucially, in situations where the agent's hypothesis space covers the number of available options (e.g. the agent thinks that the object was hidden either in box A or B and there are only two boxes), the representation of the other's ignorance and the representation of the alternatives the other likely considers yields exactly the same predictions (equal attention allocated to all options being present). Therefore, we also included locations both the observer and the agent considered to be an 'impossible' hiding place for the object, i.e. locations to which participants should not attend if they indeed represented the other's hypothesis space and not just the fact that he is ignorant ('*does not know where the object has hidden'*). To ensure that observers clearly understand and can keep in mind which options constitute the hypothesis space of the agent, 'possible' and 'impossible' meant physical possibility, that is it was determined by the physical properties of the object and the hiding places (e.g. the diameter of the hiding object, that unambiguously determined, whether the object could or could not actually fit in the respective hiding place).

To this end, we created animations in which a human avatar did not witness the hiding of a self-propelled object, that could fit in only two out of the four presented boxes, and tested whether participants represented the hypothesis space of the avatar (and directed their attention to these two possible locations) via measuring their sensitivity to subtle changes (colour change of a dot placed at the bottom of the boxes) at the three different types locations: (i) at the 'impossible' ones, where both the participant and the other agent knew the object could not hide, given its physical properties, (ii) at the object's actual hiding location, and (iii) at the other 'possible' location, where it *could have hidden,*

and which was represented as a potential hiding place by the agent. True belief trials differed from the underspecified belief trials only in that the human avatar saw the hiding of the object before turning away, thus, just like participants, he did not have any reason to sustain two alternatives. Using this paradigm, we ran four experiments. In Experiment 1a and 1b and in Experiment 3, the agent's belief was completely task-irrelevant. In Experiment 2 participants were instructed to track the agent's belief, and occasionally, upon receiving a prompt, they had to indicate which locations the agent considers 'possible. In all experiments, we hypothesized that, if participants encode the content of the other agent's hypothesis space, then, on those trials where the agent was absent during the hiding event, thus could think that the object is '*either in box A or in box B'*, upon return, they should be faster in detecting (and miss fewer) changes, not only at the actual location of the ball (i.e. the location matching their own representation of the state of the world) but also at the location that was empty but could be (and presumably was) represented as 'a potential hiding place' by the other agent, compared to changes at locations where hiding was impossible. In contrast, on trials where the other agent knew in which box the object hid, they either should not demonstrate such an attentional bias, as in those cases the other agent could eliminate one of the represented alternatives or, even if present, for example, due to an inability to fully inhibit the location that was a possible hiding place initially, the magnitude of the bias should be smaller. Importantly, if they represent such contents spontaneously, the effect should be present also when the agent's belief is not relevant for the task they perform and even if the agent is not expected to act on the object (Experiment 1a, 1b, 3a and 3b) and not only when they have to monitor the content of his belief (Experiment 2).

3.2. Experiment 1a

Experiment 1a tested whether human adults track another agent's hypotheses unintentionally and without being aware of doing so. Participants saw animations with four boxes and a ball, in which first the ball hid in one of two large boxes it could fit in, with an agent either witnessing the hiding or not, then a change happened at one of the four locations. Participants' task was to simply indicate where the change occurred. They only had to pay attention to when the other agent turned away from the scene, but not otherwise, that is, the other agent's belief was completely task-irrelevant. We hypothesized that, if participants spontaneously represent the hypotheses of the other agent, then, on those trials where the agent does not know in which of the two large boxes the ball is hiding, they should be faster to detect (and should miss fewer) changes not only at ball's the actual location, (compared to changes at locations where hiding was impossible) but also at the location the other

agent considers 'a possible hiding place' for the ball. No such bias should be present on those trials where the other agent has the same knowledge they have (or this bias towards the possible location should be less strong).

3.2.1. Methods

3.2.1.1 Participants

Participants were 26 university students (M_{age}= 22.81, SD_{age}=0.64, 10 males), recruited via a student job agency and the university's research participation system (the SONA systems). All of them were right-handed and had normal or corrected-to-normal vision as well as normal hearing. The study was approved by the EPKEB United Ethical Committee, Hungary; participants signed informed consent prior to the experiment and received monetary compensation or gift vouchers for their participation (equivalent to approximately 5 Euros).

3.2.1.2. Stimuli and apparatus

Stimuli consisted of animated videos of two types: familiarization (37.5 sec long) and test videos (23.63 sec long). Each video depicted a central human avatar, a blue ball-like agent with eyes ('ball'), extending 3.07° in diameter, and four different boxes, in two shapes, each one with a lid (see **Figure 3.1a** for an example). We decided to use a ball-like agent instead of an object, to be able to avoid all the possible issues that could have arisen from having another agent being present (who performs the hiding) with a belief content that either matches or not that of the human avatar and the participant. Two boxes were positioned on the right and two on the left side of the scene, with the central part being left empty (to avoid a potential central attention bias). Boxes had the same colour and were of the same height (each extending 3.15° vertically) but different widths, resulting in two 'large' boxes (extending 3.15° in width) in which the agent could and two 'small' ones (extending 2.09° in width), in which the ball-like agent could not fit in ('impossible' locations). They were aligned horizontally and arranged such that the two large ones, that the agent could consider a potential hiding place ('possible' locations), were always positioned on the opposite sides, either closer to the centre ('centrally') or further away from it ('peripherally'). The aim of this manipulation was to rule out the possibility that a potential attentional bias towards an empty but 'possible' location is merely the result of its proximity to the ball's actual hiding place. Different combinations of the boxes resulted in four kinds of scene

arrangements (see **Figure 3.1c**). Each box had a small grey disc displayed on the front (extending 0.32° in diameter), one of which turned red in the change videos. The discs were of the same size, were aligned horizontally, and positioned centrally, but close to the lower part of the boxes.

Familiarization videos started with the presentation of the ball in the middle of the scene, facing the participant. Following this, the ball moved either to the left or the right and started to jump on the top of the boxes, one after the other, returning to the middle, into its original position, after finished. After the ball jumped on the boxes the lid opened, and the ball fell into the boxes that were large enough but not in the other two, the openings of which were too narrow. Importantly, the human avatar was present throughout these events thus, participants could think that he had the same knowledge, regarding what is a 'possible' and an 'impossible' hiding location, as they themselves had. There were altogether eight different familiarization videos (the four arrangements, presented once with the ball starting from the left, once from the right), each of which had two versions, in which either one or the other small box opened partially[12].

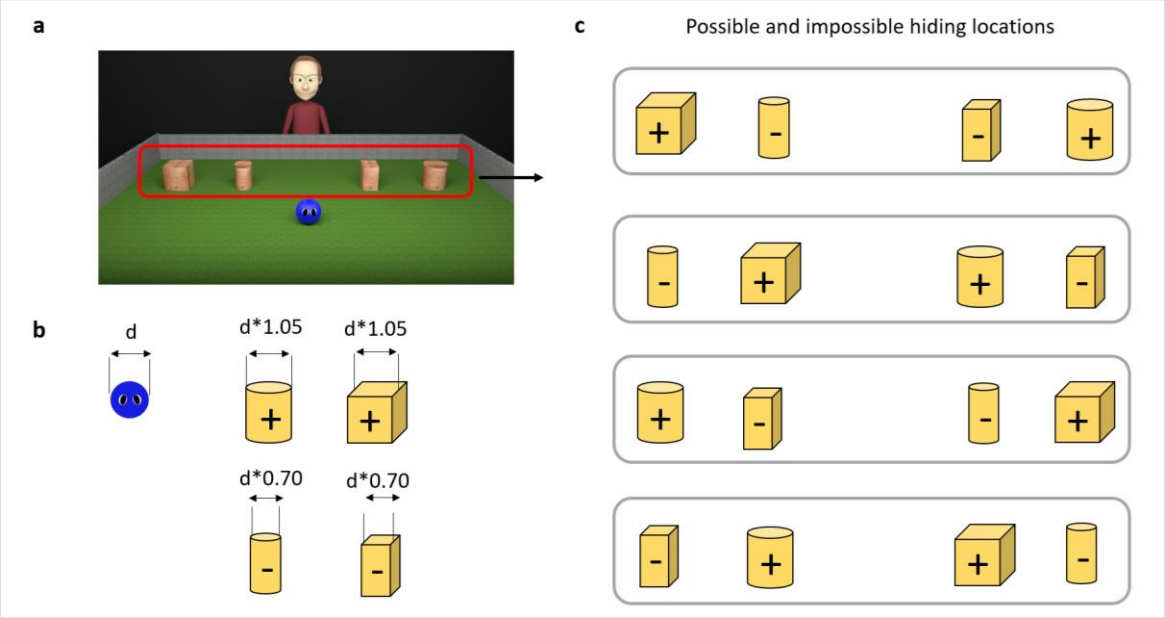


Figure 3.1. (a) An example of the scenes used. (b) The ratio of the ball and the boxes. (c) The four scene arrangements used in the familiarization and test videos. The + and the – sign denotes the possible and impossible hiding locations, respectively. The large boxes were potential hiding places for the ball, as it could fit in those ('possible' hiding locations), the smaller ones constituted 'impossible' hiding locations. The two possible locations were always positioned on the opposite sides.

[12] In specific, the lid of one of the small boxes did not open fully, making hiding impossible even for a smaller object. We planned to use this manipulation in a subsequent study, presenting both a small and a large ball, but abandoned these plans later.

The events presented in the test stimuli had two main parts: (1) a 'belief induction' (19.63 sec) and (2) an 'outcome' part (4 sec). Belief induction videos could either depict a true belief scenario (TB videos) or underspecified belief scenario (UB videos), depending on which events the human avatar witnessed. Varying the arrangement of the boxes, the ball's hiding location and the direction in which the ball started to move first in the hesitation phase resulted in sixteen versions of each.

Each 'belief induction' video had three phases, such that the first and the last phase was physically identical in the two belief conditions. They started with a 7.33 sec long 'hesitation' phase, in which the ball moved in one then in the other direction, before returning to the middle. The purpose of this phase was to highlight the uncertainty regarding where the ball 'wants to' hide, possibly increasing the likelihood of forming and attributing a belief with a disjunctive content. The 'hesitation' phase was followed by a 10.34 sec long 'belief attribution' phase, during which the ball jumped into one of the large boxes, with the human avatar either facing the scene (and then turning away) or with its back towards it. Videos ended with the avatar slowly turning back, from the 17.67 sec ('return' phase). In each video, the human avatar's turn was prompted by a telephone ring (1.4s). The two belief conditions differed only in the timing of this telephone ring and the subsequent turn. In the UB videos, the telephone started to ring at 9.16 sec, after the ball started to move towards its final location but well before it would have reached the hiding box. More specifically, it started when the ball reached the same spatial location where it turned back and started to move in the other direction in the hesitation phase, such that when the agent turned away, he could not know for sure how the ball movement would continue. In the TB condition, the telephone started to ring after the ball reached the hiding location and finished jumping in the box, at 13.49 sec. Apart from this difference, the two conditions were tightly matched with respect to the timing of the critical events: the ball's first move, the time when it reached the hiding box, finished jumping, and the moment when the agent started to turn back (for details of the timing see **Figure 3.2a**).

Outcome videos had two types, depending on the type of the presented trial, which could be either 'experimental' or 'catch' videos. Catch videos were included to check participants' attention. On the experimental trials, outcome videos started with presenting the last frame of the belief induction videos for 2 sec, after which one of the discs turned red for 750 ms. Following this, it turned back to grey again, and after 1.25 sec the video ended. Each location of change was paired with each arrangement, resulting in 16 different 'change-outcome' videos altogether. Importantly, the 'outcome' was physically identical for the TB and UB videos. 'Catch' trials differed from experimental trials in that, instead of a change, participants heard a high-pitch tone after 2 sec, which lasted for 300 ms. 'No-change-outcome' videos had four versions, one per arrangement.

The animations were generated by Maya 2019 3D software, exported as QuickTime movies, and were presented, with a screen resolution of 1280 x 720, using Psyscope B77 software

(http://psy.cns.sissa.it/), on a 19-inch LG screen (having a screen resolution of 1440 x 900). Stimuli were displayed on a plain black background. Responses were recorded by an ioLab Response box, the top of which was partially covered with a piece of paper so that only the buttons used for responding (the leftmost and the rightmost two buttons) were visible.

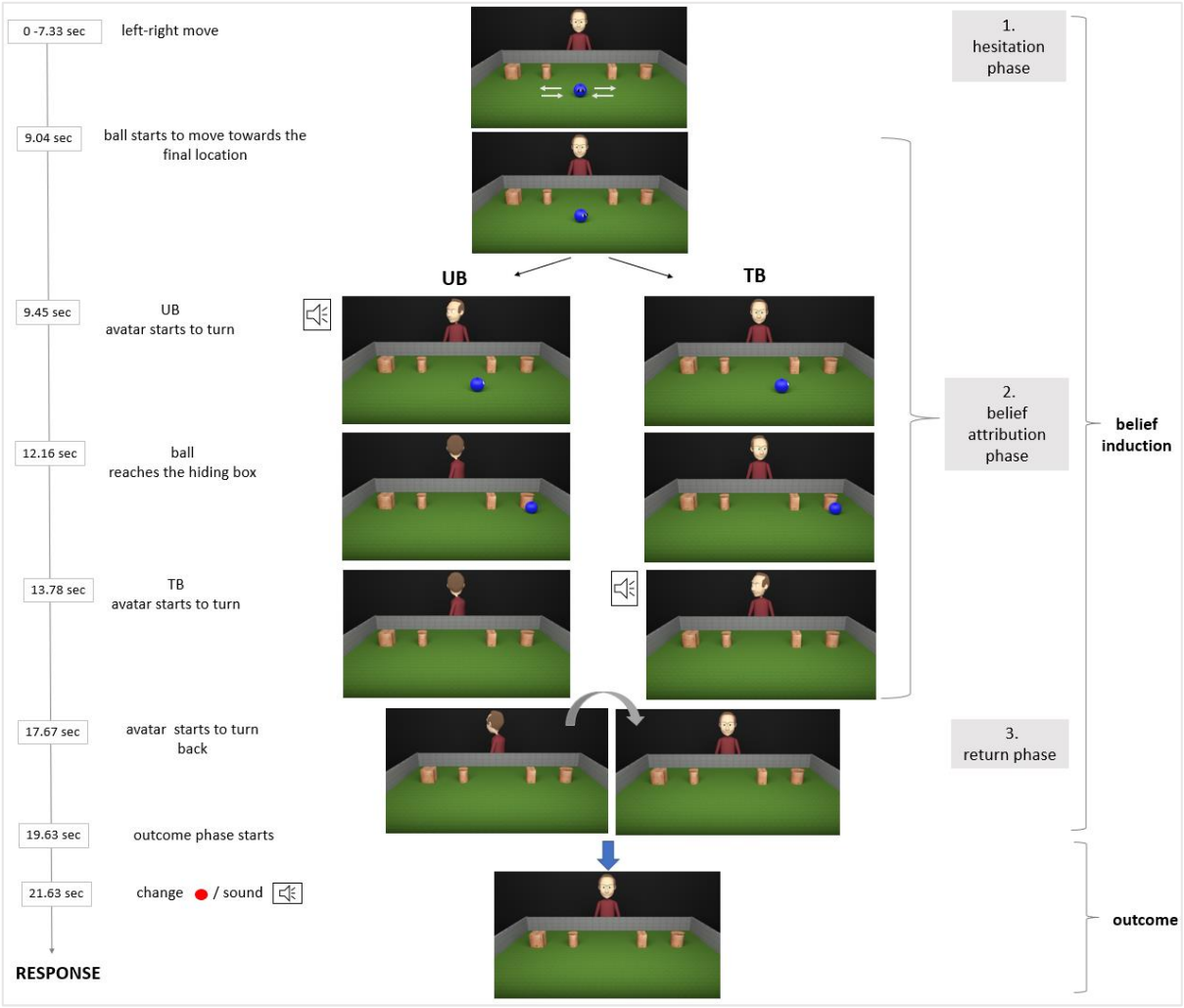


Figure 3.2. (a) Trial structure of the test videos with the timing of the critical events. Test videos were made up of two parts: the 'belief induction' and the 'outcome'. Belief induction videos had three phases and could be either underspecified (UB) or true belief (TB) videos. The two differed only in the order of the events in the middle, 'belief attribution' phase, specifically in the timing of the human avatar's turn, which defined which events he witnessed. The outcome consisted of a change in the colour of one of the grey dots at the front of the boxes: in specific, one of them turned into red for 750 ms. Participants' task was to indicate the location of change by button press. Catch trials, included to ensure participants' attention to the events, involved no change: participants heard a sound instead and they had to indicate where the ball hid.

3.2.1.3 Procedure

Participants were tested individually, in a dimly lit room, at an approximately 50 cm viewing distance from the screen. The experiment started with a familiarization phase, during which they watched eight familiarization videos), in a randomized order, with no specific instruction provided. This was followed by written and oral instructions and a short practice with oral feedback, after which the experimenter left, and the test phase started.

The test phase consisted of experimental and catch trials, which differed essentially in the outcome: while it was a 'change' video for the experimental trials, catch trials had a 'no-change' video as the second part. Each test trial started with the presentation of a central fixation cross for 500 ms. This was followed by the test video, which remained on the screen for a maximum of 2000 ms after the start of the change/sound or until a response was given. Trials were followed by a 500 ms ITI during which the screen was blank.

On experimental trials, participants' task was to indicate on which box did the disc turned red, as fast as possible, by pressing the button corresponding to the given location with their right hand. On the catch trials, i.e. in case they heard the sound, they had to indicate in which box the ball jumped in, in the same way. Catch trials were included to ensure that participants track the events, in particular, where the ball has hidden. In addition to the main task, to guarantee that they pay attention to which events the human avatar did and did not witness, participants also had to press a fifth button (positioned on the left side of the box), when the avatar started to turn away, with their left hand. Failure to respond within 2 sec (measured from 330 ms before the phone started to ring[13]) was considered as an index of potentially missing this critical event. Participants were instructed to keep their right hand at a fixed position, at an equal distance from the four buttons used for indicating the location of change, and their left hand on the button used for checking their attention. The button box was positioned such that there was a one-to-one correspondence between the location of the boxes on the screen and the location of the corresponding buttons on the button box itself.

Depending on the ball's hiding location and the location of the change, experimental trials could belong to one of the following four experimental conditions or location types: actual, possible, impossible-rectangular shaped (impossible_R) and impossible-cylinder shaped (impossible_C), where actual refers to the ball's hiding location, possible to the other (empty) large box in the scene, where the ball could have hidden and impossible to the small boxes where the ball could not hide. Combined with the two belief conditions (true / underspecified), this resulted in eight different trialtypes.

[13] This corresponded to the timepoint when the ball became invisible in the TB trials. Such a time window was set to be able to include premature responses in the analysis.

Following two practice trials (one 'change' and one 'no-change' or 'catch' trial, repeated if necessary), participants received four blocks of 40 trials, made up of 32 experimental and 8 catch trials, that were randomly intermixed with the experimental trials. Each block contained four experimental trials per trialtype, and four catch trials per belief type. The arrangement of the boxes, the direction of the ball's first move during the hesitation phase ('hesitation type') and the location of change was counterbalanced within blocks - i.e. both the position and the identity of the box on which the change occurred -, as well as the location of the hiding, with the first three factors also counterbalanced within the two belief conditions. The order of the trials within the blocks was pseudorandomized, such that there were no more than three consecutive trials with the same belief, location type, scene arrangement, and type of hesitation and no more than two consecutive trials with the same hiding or change location.

3.2.1.4 Data analysis

Analyses focused on the responses provided during the outcome phase. Catch trials were only analysed to check whether participants managed to track the events and only with respect to the hit rate and the miss ratio (computing the averages separately for the true and the underspecified trials). The primary dependent measures were (1) the reaction time (RT) and (2) the number of misses on the experimental trials, with averages calculated separately for the eight trialtypes. In addition, we also analysed the number of errors, computing the means in the same way as described above. First, we compared the two 'impossible' locations along the mean RTs, to investigate whether differences in surface features had any effect on how much attention was allocated to the two types of small boxes. If the difference was not significant, we collapsed the trials over these two trialtypes, resulting in three main location types: actual, possible and impossible.

Participants whose miss ratio was above 0.30 on either of the six main location types, whose hit rate was below 0.70 on either the true or underspecified belief catch trials, and those who failed to perform the 'attention check' on >30% of the trials belonging to any of the six main location types were not included in the statistical analyses, as such results were considered to indicate lack of motivation or inattentiveness. Trials in which the wrong button or no button was pressed and reaction times more than two standard deviations from the condition means of each participant (calculated separately for the two belief conditions) were considered invalid and were excluded from the RT analyses (for trial exclusion rates see: Supplementary Materials **S3.1**). Due to the non-normal distribution of the data, dependent variables were either log-transformed (RT data), with log-transformation carried out at trial level, or were analysed by nonparametric tests (miss ratio). Log-transformed reaction times were

investigated by a 2x3 repeated-measures ANOVA (applying Greenhause-Geisser correction whenever sphericity assumptions were not met), with belief (true and underspecified) and location type (actual, possible and impossible) as within-subject factors, and subsequent paired-samples t-tests, run separately for the TB and the UB trials. Miss ratios were analysed first by Friedman tests, performed separately for the TB and the UB trials, to compare the three main location types, then by follow-up Wilcoxon Signed Rank Tests. To avoid unnecessary loss of power, we used the Holm's Sequential Bonferroni Procedure to adjust for multiple comparisons[14]. Adjusted p-values were calculated in R (using the 'p.adjust' function of the stats package). All tests were two-tailed with significance level set at $p<0.05$.

Our three crucial tests were: (1) the comparison of RTs and miss ratios for changes at the actual and the impossible locations in the true belief condition (baseline effect) and (2) the comparison of RTs and miss ratios for changes at the possible and impossible locations, in the underspecified as well as in the (3) true belief condition. While the first comparison tested for the validity of the paradigm, i.e. whether participants' own representation of the ball's location influenced the allocation of their attention and, via this, their sensitivity to changes at the represented location, the second and third ones tested for the actual hypothesis, i.e. that adults spontaneously represent the alternatives of another agent.

To test whether the crucial differences are present already in the first half of the experiment or develop only with time (by the second half), we also ran a 2 (time: 1st half/2nd half) x 2 (belief:-TB belief versus UB belief) x 3 (location: actual, possible, impossible) ANOVA (with appropriate follow-up tests) on the log-transformed RT data as well as Friedman tests, separately for the miss ratios obtained in the first and the second half of the task. In addition, we tested whether the actual spatial position of the boxes (central versus peripheral) had an effect on the reaction times and the number of misses. The results of these additional analyses are reported in the Supplementary Materials (Section S3.2).

Besides the analyses ran at the group level, we also computed the difference of the impossible and possible trials' mean RT, separately for the two beliefs, for each participant. Having a positive difference score on the UB trials and either (a) a negative or zero difference score on the TB trials or (b) a positive score smaller than the one obtained for the UB trials, was taken as evidence for spontaneously tracking the other's underspecified belief content, at an individual level, independent of the magnitude of the UB effect.

[14] Whenever the p-value would have exceeded 0.99 after adjustment, we report the unadjusted value.

3.2.2 Results

3.2.2.1 Reaction time analyses

Main analyses

Paired samples t-tests, run separately for the TB and UB trials, revealed no significant difference between how fast participants reacted to changes at the rectangular and the cylinder-shaped 'impossible' location (TB: $t(25)=-0.10$, $p_{unadj}= .925$, $d=0.018$; UB: $t(25)=0.14$, $p_{unadj}= .889$, $d=0.028$). Therefore, data from these two types of trials were collapsed, and all subsequent analyses were run with three location types.

2 x 3 repeated-measures ANOVA with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors, yielded a significant main effect of belief ($F(1, 25)=9.13$, $p= .006$, $\eta_p^2= .268$) as well as location type ($F(1.34, 33.68)=16.83$, $p< .001$, $\eta_p^2= .402$). Importantly, there was also a significant belief x location type interaction ($F(2, 50)=4.63$, $p= .014$, $\eta_p^2= .156$) resulting from a marked difference in how fast participants detected changes at the 'possible' (but empty) location, in the two belief conditions (see **Figure 3.3a**). Pairwise comparisons revealed that participants were the fastest in detecting changes at the actual compared to the impossible location, both on TB (actual-impossible: $t(25)=-3.09$, $p_{adj(3)}= .010$, $d=0.606$; actual-possible: $t(25)=-4.53$, $p_{adj(3)}< .001$, $d=0.888$) and UB trials (actual-impossible: $t(25)=-4.77$, $p_{adj(3)}< .001$, $d=0.93$; actual-possible: $t(25)=-2.35$, $p_{adj(3)}= .027$, $d=460$). Crucially, however, in the underspecified belief condition, they were significantly faster in detecting changes not only at the actual location of the ball, compared to changes at the 'impossible' locations, but also at the location the other agent could consider a 'potential' hiding place in the given situation (possible-impossible: $t(25)=-2.70$, $p_{adj(3)}= .024$, $d=0.530$)[15], suggesting that they encoded both alternatives presumably represented by the other. Surprisingly, a reverse pattern emerged in the true belief condition: participants were significantly slower in detecting changes at the possible compared to the impossible locations ($t(25)=2.31$, $p_{adj(3)}= .030$, $d=0.451$) indicating inhibition of this location. Importantly, this difference in the RT pattern – shorter RTs in the UB and longer RTs in the TB condition for changes at the 'possible' location - was not due to participants paying less attention to where the ball hid (TB-UB actual: $t(24)=1.43$, $p_{adj(3)}= .332$, $d=0.280$) or being significantly slower in detecting changes at the impossible locations, on the UB compared to the TB trials (TB-UB impossible: $t(24)=-0.60$, $p_{adj(3)}=0.555$, $d=0.117$).

[15] The number in the subscript the refers to the number of comparisons taken into account when calculating the adjusted p-value.

The difference scores (impossible-possible RT), computed for the UB and TB trials, provided evidence for spontaneous third-person representation of alternatives in case of 16 (62%) participants. There were only 7 participants who did not show the predicted effect on the UB trials (difference score≤0).

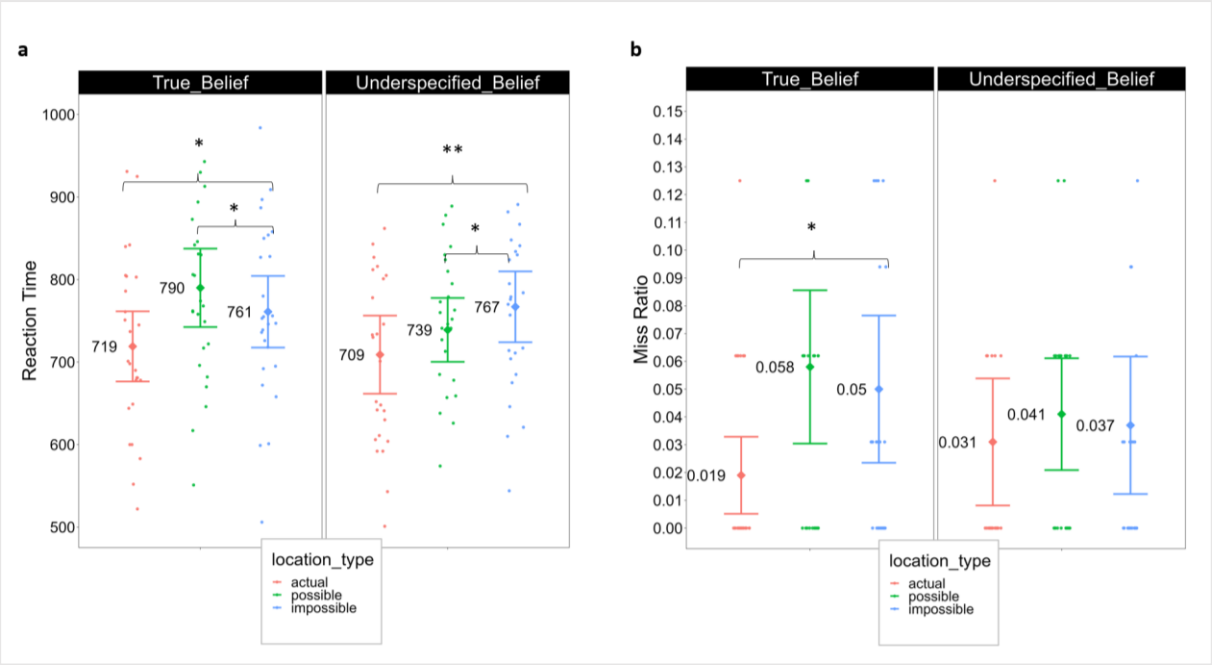


Figure 3.3. (a) Mean reaction time and (b) miss ratio in the true and underspecified belief conditions, for changes occurring at the actual, possible and impossible locations in Experiment 1a. Error bars represent 95% CI, dots show the individual means. Note: Only the significance levels of the two target comparisons are indicated (actual versus impossible and possible versus impossible) on the figures. For the other comparisons see the main text. The statistical tests for the RT differences were run on the log-transformed RT data.
+: p<0.1; *: p<0.05, **: p<0.01.

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

A 2 x 2 x 3 repeated measures ANOVA with time (first vs second half of the experiment), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors yielded a significant main effect of time ($F(1, 25)=32.24$, $p< .001$, $\eta_p^2= .563$), resulting from generally shorter reaction times in the second half of the task, as well as significant main effect of belief ($F(1, 25)=8.53$, $p= .009$, $\eta_p^2= .243$), location type ($F(1.33, 33.30)=15.89$, $p< .001$, $\eta_p^2= .389$) and a significant belief x location type interaction ($F(1.90, 47.41)=3.70$, $p= .034$, $\eta_p^2= 129$). In addition, there was also a tendency level time x location type ($F(1.68, 42.08)=0.26$, $p= .094$, $\eta_p^2= .094$), resulting from the fact that the general drop in RT was more pronounced for changes at the possible than at the other two types of locations. Despite this had an opposite effect on difference the between the possible and the impossible location's RT on the TB and the UB trials (with the difference almost disappearing on the TB and becoming more

pronounced on the UB trials, see Supplementary Materials **Figure S3.9a** and **S3.9b**), neither the time x belief ($F(1, 25)=0.82$, $p= .375$, $\eta_p^2= .032$) nor the time x belief x location type interaction was significant ($F(2, 50)=0.60$, $p= .555$, $\eta_p^2= .023$). Importantly, the difference in how fast participants reacted to changes at the 'possible' versus the 'impossible' locations on the underspecified belief trials was present from the very beginning of the experiment and remained there until the end of the task.

3.2.2.2 Hit rate and miss ratio analyses

Main analyses

The hit rate was at ceiling (all averages>0.99), with no significant difference between the three location type, on either of the two types of belief trials (TB: $\chi^2(2)=0.50$, $p= .779$, Kendall's $W=0.010$; UB: $\chi^2(2)=2.00$, $p= .368$, Kendall's $W=0.038$). The number of misses was also very low, although higher than the number of errors, with a large individual variation in all experimental conditions (see **Figure 3.3b**). Friedman test, run on the TB trials, revealed a significant difference between the three types of locations, with respect to how many times participants failed to detect the changes at those ($\chi^2(2)=8.60$, $p= .014$, Kendall's $W=0.165$). Post hoc Wilcoxon Signed Rank Tests indicated that participants had significantly more misses both on the 'possible' ($Z=-2.40$, $p_{adj(3)}= .033$, $r=0.471$) and the 'impossible' trials ($Z=-2.55$, $p_{adj(3)}= .033$, $r=0.500$) compared to those trials where the change occurred at the actual location of the ball, suggesting that they allocated less attention to the empty boxes when the other agent knew where the ball was. There was no such difference between the three location types on the UB trials ($\chi^2(2)=1.24$, $p= .538$, Kendall's $W=0.024$), indicating a less marked attentional bias towards the object's actual location when the human avatar was uncertain about where it hid. Despite the different patterns, the TB and the UB trials did not differ significantly with respect to the number of misses at any of the three types of locations (all Zs <1.56, p_{unadj}s >.190, r<0.306).

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

To investigate whether the observed pattern of results was stable over time, we ran a series of Friedman tests, comparing the three location types separately for the first and the second half of the

trials and for the two belief types. Regarding the TB trials, although the number of misses was higher for changes at either one or the other empty location compared to changes at the actual location of the ball, from the beginning of the experiment (see Supplementary Materials **Figure S3.9c** and **S3.9d)**, the difference between the location types became significant only by the second half of the task ($\chi^2(2)=7.42$, $p= .030$, Kendall's $W=0.135$; actual-impossible: $Z=-2.56$, $p_{adj(3)}= .030$, $r=0.502$; actual-possible: $Z=-2.14$, $p_{adj(3)}= .066$, $r=0.420$). As for the UB trials, the initially nonsignificant difference between the three location types (1st half: $\chi^2(2)=2.78$, $p= .249$, Kendall's $W=0.053$) became marginally significant with time (2nd half: $\chi^2(2)=5.07$, $p= .079$, Kendall's $W=0.97$), due to a marked decrease in the number of trials on which participants failed to detect changes at the 'possible' location by the second half of the task, which resulted in a more pronounced although nonsignificant difference between the possible and the impossible location ($Z=-1.93$, $p_{adj(3)}= .162$, $r=0.379$).

3.2.2.3 Catch trials

As for the catch trials, the hit rate was at ceiling (average>0.98 for both the TB and UB trials) and the number of misses was very low (TB: $M=0.019$ $SD=0.029$; UB: $M=0.019$ $SD=0.049$), just like in the experimental trials, indicating no difficulty to track the location of the ball and remember this information.

3.2.3 Discussion

The results of Experiment 1a indicated that participants represented both alternatives the other agent, who did not witness the hiding, presumably did, and maintained this representation in their working memory (or recalled it after the agent turned back). They were faster in detecting changes not only at the actual location of the ball but also at the other, 'possible' location, when the other was uncertain about where the ball has hidden, compared to those locations where hiding was impossible, despite they could clearly track the ball's location (as indicated by, for example, their high level of performance on the catch trials), therefore had no reason to sustain the information about the other alternative. By the second half of the task the attentional bias, towards the possible location, also emerged in the number of misses. In line with these findings, while there was a strong attentional bias towards the

actual location of the ball on the true belief trials, that was reflected not only in the reaction times but also in the number of trials where participants failed to detect changes, such a bias was not present in the number of misses when the human avatar was uncertain about the location of the ball (though it emerged in the reaction time pattern).

Surprisingly, when the other agent co-witnessed the hiding, participants' performance was worse for changes at the empty location where the ball *could have* hidden, both in terms of reaction time and the number of misses, suggesting that they might have inhibited ('negated') this location when both they themselves and the other agent knew that the ball was not there. That is, instead of simply representing '*the ball is at location A'* they set up the representation '*the ball is at location A and not at location B*' (and may or may not have attributed this content to the other agent). Importantly, if the observed reaction time pattern indeed reflects inhibition of the particular location this suggests that participants represented both alternatives at some point before the hiding would have taken place (either during the hesitation phase or after it had finished), from first-person perspective. Hence, these results may provide evidence for the spontaneous elimination of one element of a previously represented disjunction, via negation, on the basis of the observed events.

Given our unpredicted findings on the true belief trials, on the one hand, and the current replication crisis in the field of ToM on the other, we decided to run a direct replication of our first experiment as a next step, to make sure that we have a solid ground for moving on to investigate further aspects' human adults' capacity to ascribe underspecified beliefs to other agents and for making specific claims about the format of the attributed content.

3.3. Experiment 1b

Experiment 1b was a direct replication of Experiment 1a, with a minor modification in the counterbalancing and a larger sample size.

3.3.1 Methods

3.3.1.1 Participants

Participants were 33 university students (M_{age}=22.18, SD_{age}=2.72, 15 males), recruited via a student job agency and the university's research participation system (the SONA systems). All of them were right-handed and had normal or corrected-to-normal vision as well as normal hearing. Additional three students were tested but their data was excluded because participants did not meet the inclusion criteria (miss ratio was >0.30 in multiple conditions N=2; or failed to perform the 'attention check' on >30% of the trials in more than one condition: N=1).

The target sample size was determined by power analysis (using G*Power 3.1) based on the effect size obtained in Experiment 1a in the comparison of RTs for changes at the possible versus the impossible locations on the UB trials (d=0.530) for a two-sided paired-samples t-test with alpha set at 0.05. A sample size of N=30 was estimated to provide an adequate 80% statistical power. Calculating with a maximum of 20% exclusion rate resulted in testing N=36 participants.

The study was approved by the EPKEB United Ethical Committee, Hungary; participants signed informed consent prior to the experiment and were compensated the same way as in Experiment 1a.

3.3.1.2 Stimuli, apparatus, procedure, and data analysis

The stimuli, the apparatus, and the procedure used were the same as in Experiment 1a, with one minor difference: the hiding location of the ball was also counterbalanced within belief and block.

The data was analysed in the same way as in Experiment 1a.

3.3.2 Results

3.3.2.1 Reaction time analyses

Main analyses

There was no significant difference between how fast participants reacted to changes at the two types of (rectangular and cylinder-shaped) 'impossible' locations (TB: $t(32)=-0.94$, $p_{unadj}= .353$, $d=0.164$; UB: $t(32)=0.65$, $p_{unadj}= .519$, $d=0.107$). Therefore, data from these two types of trials were collapsed, and all subsequent analyses were run with three location types.

A 2 x 3 repeated-measures ANOVA with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors yielded a significant main effect of location type ($F(2,64)=11.85$, $p< .001$, $\eta_p^2= .270$), resulting from the fact that participants were generally faster to detect changes at the actual location of the ball than at either of the two other types of empty locations, but no significant main effect of belief ($F(1,32)=1.85$, $p= .184$, $\eta_p^2= .055$) or belief x location type interaction ($F(2,64)=0.15$, $p= .860$, $\eta_p^2= .005$), reflecting the similar pattern in the two belief conditions. Crucially, as can be seen on **Figure 3.4a**, in this experiment participants' reaction times were highly similar for the changes at the possible and the impossible locations, on both the true and the underspecified belief trials. That is, in this sample, neither the predicted effect (lower RTs on the possible trials, in the underspecified belief condition) nor the surprising, reverse effect (higher RTs on the possible trials, in the true belief condition) was present (UB possible-impossible: $t(32)=-0.99$, $p_{adj(3)}= .331$, $d=0.172$; TB possible-impossible: $t(32)=-1.08$, $p_{adj(3)}= .236$, $d=0.188$). Pairwise comparisons revealed significant differences only between the actual and the other two locations on both type of belief trials (TB - actual-impossible: $t(32)=-3.83$, $p_{adj(3)}= .003$, $d=0.667$; actual-possible: $t(32)=-2.79$, $p_{adj(3)}= .018$, $d=0.486$; UB - actual-impossible: $t(32)=-3.59$, $p_{adj(3)}= .003$, $d=0.625$) with the actual versus possible difference being only marginally significant on the UB trials ($t(32)=-2.25$, $p_{adj(3)}= .064$, $d=0.392$).

There were only 12 (36%) participants whose difference scores (impossible-possible RT) were in line with our predictions, i.e. for whom there was some evidence that they may have tracked the hypotheses of the other agent. The variation in terms of the magnitude of the differences, was, however, very high even in the subgroup showing the predicted pattern (see Supplementary Materials **Figure S3.7b**). Crucially, almost half of the participants (N=16, 48%) did not show the predicted effect (UB trials impossible-possible RT>0 ms).

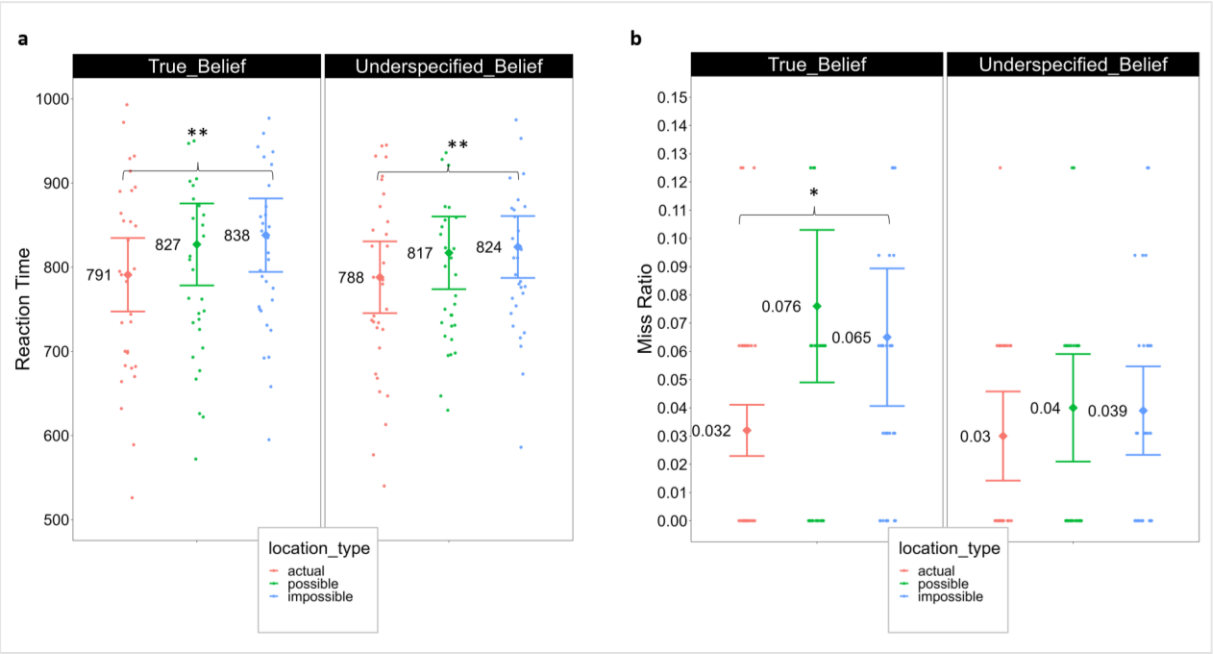


Figure 3.4. (a) Mean reaction time and (b) miss ratio in the true and underspecified belief conditions, for changes occurring at the actual, possible and impossible locations in Experiment 1b (direct replication of Experiment 1a). Error bars represent 95% CI, dots show the individual means. Note: Only the significance levels of the two target comparisons are indicated (actual versus impossible and possible versus impossible) on the figures. For the other comparisons see the main text. The statistical tests for the RT differences were run on the log-transformed RT data. +: p<0.1; *: p<0.05, **: p<0.01

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

A 2 x 2 x 3 repeated measures ANOVA with time (1st versus 2nd half), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors revealed a significant main effect of time ($F(1, 32)=44.06$, $p< .001$, $\eta_p^2= .579$), reflecting the marked drop in participants' RTs by the second half of the task. There was also a significant main effect of location type ($F(1.75, 55.91)=12.24$, $p< .001$, $\eta_p^2= .277$) but no significant main effect of belief ($F(1, 32)=1.42$, $p= .243$, $\eta_p^2= .042$) or a belief x location type interaction ($F(2, 64)=0.12$, $p= .887$, $\eta_p^2= .004$). Importantly, despite participants' reaction time pattern differed in the first and the second half of the experiment on the TB trials, due to the fact that participants were faster to detect changes not only at the actual but also at the possible (but empty) location compared to the impossible locations on these trials in the first (but not in the second) half of the task, and that such difference was not present on the UB trials, none of the interactions with time were significant (the time x belief: $F(1,32)=0.37$, $p= .545$, $\eta_p^2= .012$; time x location type: $F(2, 64)=1.42$, $p= .249$, $\eta_p^2= .043$; time x belief x location type: $F(1.71, 54.79)=0.49$, $p= .585$, $\eta_p^2= .015$).

3.3.2.2 Hit rate and miss ratio analyses

Main analyses

The hit rate was at ceiling (all averages>0.99), with no significant difference between the three location types, on either of the TB (χ^2(2)=4.00, p= .135, Kendall's W=0.061) or the UB trials (χ^2(2)=1.00, p= .607, Kendall's W=0.015).

Friedman test, performed on the TB trials, indicated a significant difference between the three location types in terms of the number of misses (χ^2(2)=6.47, p= .039, Kendall's W=0.098). Post hoc Wilcoxon Signed Rank Tests revealed that participants failed to detect the change significantly more often when it happened at the 'possible' or the 'impossible' locations (see **Figure 3.4b**) than when it occurred at the actual location of the ball (possible-actual: Z=-2.78, $p_{adj(3)}$= .015, r=0.484; actual-impossible: Z=-2.35, $p_{adj(3)}$= .038, r=0.409), suggesting that they directed less attention to the empty boxes when the other agent knew where the ball was. There was no such difference between the three types of locations on the UB trials (χ^2(2)=1.95, p= .377, Kendall's W=0.030), indicating that the attentional bias towards the actual location of the object (present in the RTs) was somewhat less marked when the human avatar was uncertain about where it has hidden. This diminished attentional bias was also indicated by the significantly lower number of misses on the possible and the impossible trials in the UB compared to the TB condition (possible: Z=-2.28, $p_{adj(3)}$= .046, r=0.397; impossible: Z=-2.64, $p_{adj(3)}$= .024, r=0.460).

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

As for the TB trials, although the number of misses was much lower for changes at the actual location of the ball than for changes at either of the two empty locations throughout the task, the difference between the three locations, specifically the actual and the other two location types was significant only in the first (χ^2(2)=9.11, p= .011, Kendall's W=0.138; actual—impossible: Z=-2.92, $p_{adj(3)}$= .012, r=0.509; actual-possible: Z=-2.50, $p_{adj(3)}$= .024, r=0.436) but not in the second half of the task (χ^2(2)=2.70, p= .259, Kendall's W=0.041; follow-up Wilcoxon Signed Rank Tests: all Zs < 0.95, all p_{unadj}s > .411). In contrast, while there was no difference between the three types of locations on the UB trials in the first half of the task (χ^2(2)=0.22, p= .865, Kendall's W=0.008), a modest attentional bias emerged towards the actual location by the second half (χ^2(2)=4.93, p= .085, Kendall's W=0.075): participants

missed fewer changes at the actual than at the impossible locations, with difference being a marginally significant (Z=-2.32, $p_{adj(3)}$= .060, r=0.404).

3.3.2.3 Catch trials

With respect to the catch trials, the hit rate was at ceiling (average>0.99 in all conditions) and the number of misses was very low (TB: M =0.010, SD=0.015; UB: M =0.004, SD=0.014), indicating that participants could easily track the location of the ball and recall this information when needed.

3.3.2.4 Comparison of Experiment 1a and Experiment 1b

Reaction times

To investigate whether the two experiments differed significantly in terms of how fast participants detected changes at the three types of locations, in the two belief conditions, we performed a 2x2x3 mixed ANOVA, on the log-transformed RT data, with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject and experiment as between-subject factor. The analysis yielded a significant main effect of belief (F(1 ,57)=9.98, p= .003, η_p^2= .149), location type (F(1.60, 90.93)=28.16, p< .001, η_p^2= .331) and a tendency level belief x location type interaction (F(2, 114)=2.68, p= .073, η_p^2= .045). Importantly, there was also a significant main effect of experiment (F(1, 57)=4.85, p= .032, η_p^2= .078): participants were much slower to detect changes in Experiment 1b than in Experiment 1a, at all three location types and on both types of belief trials, as reflected by the lack of significant experiment x location type (F(1.60, 90.93)=1.13, p= .316, η_p^2= .019) or experiment x belief (F(1, 57)=1.83, p= .182, η_p^2= .031) interaction. There was also a tendency level experiment x belief x location type interaction, resulting from the different pattern of findings in the two experiments (F(1.99, 113.16)=2.64, p= .076, η_p^2= .044), i.e. the nonreplication of the original results in both the true and the underspecified belief condition.

Miss ratios

Comparison of the two experiments along the miss ratios revealed no significant difference between the two experiments in either of the six main experimental conditions (Mann-Whitney tests: all Us>350.00, all p_{unadj}s> .203).

3.3.3 Discussion

Contrary to our expectations, despite no change in the stimuli or the procedure, Experiment 1b did not replicate the main findings of Experiment 1a: unlike in the previous experiment, participants were not faster in detecting changes at the possible but empty location compared to the impossible locations, when the agent could consider it as a potential hiding place for the object. Their reaction times indicated no differentiation of the empty boxes in terms of how much attention was allocated to them. In line with this, the ratio of participants for whom there was some evidence for representing two alternatives, from third-person perspective, was only roughly half of those for whom we had such evidence in the previous experiment.

There was also no sign of eliminating the other alternative, via inhibitory processes. Representing 'not B', should have resulted in longer RTs in the possible compared to the impossible condition, when the agent saw the hiding, but this was not the case. In fact, there was no sign of any other difference between the two conditions on the true belief trials, indicating that participants may not have even represented the two alternatives from first-person perspective (at the beginning of the trials). Importantly, this does not mean that the content of participants' own representations did not affect their attention, and via this, their reaction times in this experiment. Just like in Experiment 1a, participants in Experiment 1b were faster to detect changes at the actual location of the ball on both the true and the underspecified belief trials, providing further evidence that the paradigm is suitable for measuring of what is being represented in a given context.

Interestingly, in terms of miss ratios, the pattern of findings was the same in the two experiments: there was an attentional bias towards the actual location of the ball when the other agent knew where it has hidden on the one hand, and no specific bias towards any of the locations when he was uncertain about the location of the object on the other (at least in the first half of the experiment). This finding suggests that the content of the other's belief, more specifically, the fact whether the agent had the same knowledge as the participant or lacked full knowledge about the state of affairs, did have some effect on participants' spatial attention and, via this, their sensitivity for changes at the different locations. However, we must note that the observed pattern indicates rather the attribution of ignorance or the representation of the other's uncertainty than the spontaneous encoding of the other agent's hypotheses.

There may be several reasons why the predicted effect was not present in participants' reaction times on the underspecified belief trials. One option is that the modulatory effect of the other's spontaneously represented belief content may emerge in reaction times only if participants are fast enough, and there is no room for deliberate reasoning at the time of the response (which may result

in inhibiting the interfering content). For participants in Experiment 1b this was clearly not the case: they were surprisingly slow in all conditions, either due to motivational problems or because they found the task more difficult than participants of the previous experiment. Another option is that our reaction time measure is simply not sensitive enough to reliably capture the impact of another agent's underspecified belief content on adults' performance: even if the alternatives the other agent considers are represented by the participant, at some point of the trial, the effect is not strong enough to reliably emerge in how fast participants detect changes at the location(s) corresponding to the attributed content.

To investigate whether our paradigm can measure the effect of another agent's underspecified belief content on adults' own representation of the state of affairs, as a next step, we decided to run a version of our experiment in which we could be confident that such contents are represented by the participants. More specifically, we investigated whether the predicted effect (attentional bias toward the possible location when the agent is uncertain about the location of the object) emerges in a task where participants are instructed to track the other agent's belief regarding the location of the ball and are incidentally tested for this, by asking them to indicate the location(s) where the ball may be according to the other.

3.4. Experiment 2

Experiment 2 differed from Experiment 1b in one crucial aspect: on the 'catch' trials, i.e. upon hearing the sound, participants had to indicate the human avatar's belief, instead of the ball's location, in particular, what the agent thinks *where the ball may be*. On the experimental trials, the task was the same as in the previous experiment, hence, the measure itself remained implicit. Importantly, just like in Experiment 1a and 1b, participants did not know in advance which type of trial they will receive, therefore they had to track the agent's belief about the ball's location continuously (and they were asked to do so, to be able to perform well). Such a design allowed us to test whether a voluntarily computed underspecified belief content of another agent can involuntarily bias adults' spatial attention, and via this, influence their performance on an irrelevant task, namely how they detect changes at the locations corresponding to the content.

3.4.1 Methods

3 4.1.1 Participants

Participants were 35 university students recruited via student job agencies and the university's research participation system (M_{age}=21.43, SD_{age}=2.70, 13 males). All of them were right-handed and had normal or corrected-to-normal vision as well as normal hearing. Ten additional students were tested but their data was not analysed due to their low performance on the catch (i.e. 'explicit') trials (see the exclusion criteria below). The data of one more participant was excluded because it did not meet the inclusion criteria (<0.30 miss ratio in all conditions). The study was approved by the EPKEB United Ethical Committee, Hungary; participants signed informed consent prior to the experiment and received the same compensation for participation as in the previous two experiments.

3.4.1.2 Stimuli, apparatus, procedure

The stimuli and the apparatus were the same as in Experiment 1b. The procedure differed from the one used in the previous experiment in one crucial aspect (and some minor ones that were the consequences of this major factor): on the 'catch' trials participants had to indicate '*in which box the ball may be*, *according to the character*', upon hearing the sound (displayed instead of the change, on these trials). In both the oral and the written instruction, we empathized that they can press more than one button (one after the other) in these cases and that they have to attend to what the avatar thinks on all trials to be able to perform well. The three additional differences were the following. First, to make sure that participants understand that in case the avatar did not witness the hiding he represents two equally likely alternatives, the experimenter asked a pre-set list of leading questions in case the participant provided an incorrect response on the explicit practice trial (i.e. pressing one or no button on this trial), asking, for instance, whether the other agent has seen the hiding, pointing out that he may consider more than one alternative in case he did not and requesting the participant to identify those. Second, in case of incorrect answers, the trial could be repeated a maximum of three times, leading to a somewhat more extensive training before the test phase started than previously (two repetitions in case of N=11 and three in case of N=1 participant). Third, unlike the experimental trials, these explicit 'catch' trials ended after a fixed response period of 2500 ms following the onset of the sound, independent of whether or when the button(s) were pressed. Finally, to ensure that

participants have enough time to press both buttons on the underspecified 'catch' trials and to keep the length of the response period constant across the two trial types, test videos remained on the screen for a maximum of 2500 ms after the start of the change/sound (instead of the previous 2000 ms).

3.4.1.3 Data analysis

Besides the exclusion criteria used in the previous experiments (see Section 2.1.4), in Experiment 2 we also used two exclusion criteria based on participants' performance on the 'catch' trials, to make sure that we only include participants who explicitly represented both alternatives the other did, on the underspecified trials. In particular, we also excluded participants who either pressed (1) two buttons on ≥50% of the true belief 'catch' trials (indicating both the 'possible' and the actual location despite the agent saw the hiding, N=5) or (2) only one button on ≥50% of the underspecified 'catch' trials (N=5), as this was considered to reflect a failure to understand the task or that the agent represents two equally likely alternatives if he did not witness the hiding event, respectively.
Data from experimental trials were analysed as before. Data from 'catch' (i.e. explicit) trials was analysed only with respect to hit ratios.

3.4.2 Results

3.4.2.1 Reaction time analyses

Main analyses

Paired samples t-tests, run separately for the TB and UB trials, revealed no significant difference between how fast participants reacted to changes at one or the other type of 'impossible' location on the UB trials ($t(34)=0.50$, $p_{unadj}= .622$, $d=0.084$). There was some difference on the TB trials but it did not reach significance either ($t(34)=1.74$, $p_{unadj}= .092$, $d=0.294$). Following the practice of the previous two experiments, we collapsed data from these two types of trials and ran all subsequent analyses with three location types.

A 2 x 3 repeated-measures ANOVA with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors, yielded a significant main effect of location type ($F(2, 68)=14.64$, $p< .001$, $\eta_p^2= .301$), but no significant main effect of belief ($F(1, 34)=0.14$, $p= .713$, $\eta_p^2= .004$) or significant belief x location type interaction ($F(2, 68)=1.52$, $p= .225$, $\eta_p^2= .043$). As can be seen on **Figure 3.5a**, participants were faster to detect changes not only at the actual but also at the possible location of the ball, compared to the impossible locations, particularly on the underspecified belief trials, i.e. when the other agent could consider the empty location a 'potential' hiding place for the object (though there was some difference also on the TB trials). In line with this, pairwise comparisons revealed a significant difference between these two location types only on the UB (possible-impossible: $t(34)=-2.49$, $p_{adj(3)}= .036$, $d=0.421$) but not on the TB trials (possible-impossible: $t(34)=-0.77$, $p_{adj(3)}= .446$, $d=0.130$). There was no significant difference in how fast participants reacted to changes at the actual and the other possible (but empty) location on the UB ($t(34)=-1.46$, $p_{adj(3)}= .153$, $d=0.247$) but significant difference on the TB trials ($t(34)=-3.19$, $p_{adj(3)}= .006$, $d=0.539$), while the difference between the actual and the impossible location was significant in both cases (TB: $t(34)=-3.60$, $p_{adj(3)}= .003$, $d=0.609$; UB: $t(34)=-4.58$, $p_{adj(3)}< .001$, $d=0.773$). Importantly, the shorter RTs for changes at the possible location on trials where the agent was uncertain about the ball's location were not due to participants paying less attention to where the ball hid in these cases compared to the situation where the agent co-witnessed the hiding (TB-UB actual: $t(34)=0.86$, $p_{adj(3)}= .742$, $d=0.145$).

There were 19 (54%) participants whose reaction time pattern was in line with our predictions, i.e. for whom the difference scores, computed for the UB and TB trials, provided evidence for third-person representation of alternatives (see Supplementary Materials **Figure S3.7c**). There were 12 (34%) participants who did not show the predicted effect on the UB trials (had longer RTs for changes at the possible compared to the impossible locations).

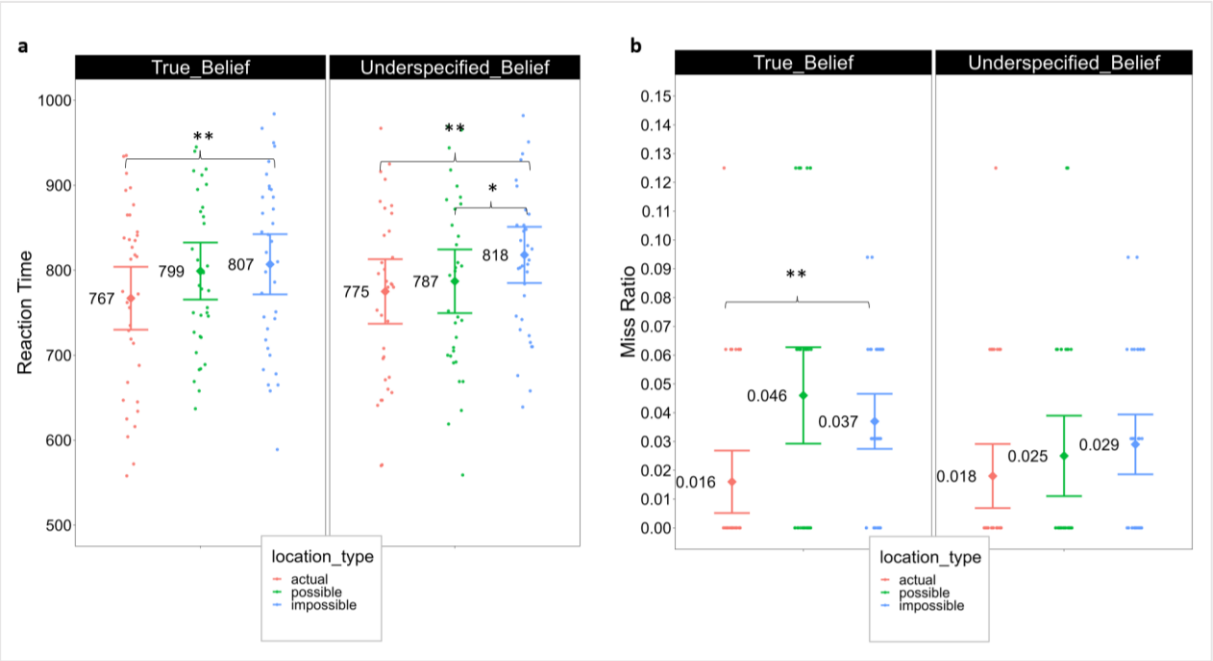


Figure 3.5. (a) Mean reaction time and (b) miss ratio in the true and underspecified belief conditions, for changes occurring at the actual, possible and impossible locations in Experiment 2 (semi-explicit version of Experiment 1b). Error bars represent 95% CI, dots show the individual means. Note: Only the significance levels of the two target comparisons are indicated (actual versus impossible and possible versus impossible) on the figures. For the other comparisons see the main text. The statistical tests for the RT differences were run on the log-transformed RT data. +: p<0.1; *: p<0.05, **: p<0.01

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

A 2 x 2 x 3 repeated measures ANOVA with time (1st versus 2nd half), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors revealed a significant main effect of time ($F(1, 34)=30.68$, $p< .001$, $\eta_p^2= .474$), resulting from the fact that participants were much faster in the second than in the first half of the task. There was also a significant main effect of location type ($F(2, 68)=15.57$, $p< .001$, $\eta_p^2= .314$) but no significant main effect of belief ($F(1,34)=0.31$, $p= .579$, $\eta_p^2= .009$) or a belief x location type interaction ($F(2, 68)=0.12$, $p= .887$, $\eta_p^2= .004$). Importantly, none of the interactions with order were significant (time x belief: $F(1, 34)=0.42$, $p= .838$, $\eta_p^2= .001$; time x location type: $F(2, 68)=0.66$, $p= .520$, $\eta_p^2= .019$; time x belief x location type: $F(2, 68)=0.629$, $p= .536$, $\eta_p^2= .018$): although its magnitude decreased with time (see Supplementary Materials **Figure S3.11a** and **b**), the predicted effect (i.e. faster reactions to changes at the possible versus the impossible locations, on the underspecified trials) was present from the beginning of the experiment and remained there until the end of the task.*3.4.2.2 Hit rate and miss ratio analyses*

Main analyses

The hit rate was at ceiling (all averages>0.99), with no significant difference between the three location types, on either of the two types of belief trials (TB: $\chi^2(2)=2.00$, $p= .368$, Kendall's $W=0.029$; UB: $\chi^2(2)=1.63$, $p= .444$, Kendall's $W=0.023$).

Friedman test, performed for the TB condition, indicated a significant difference between the three types of locations with respect to how many times participants failed to detect changes at those ($\chi^2(2)=10.84$, $p= .004$, Kendall's $W=0.155$). Post hoc Wilcoxon Signed Rank Tests revealed that participants had significantly more misses both on the 'possible' ($Z=-2.81$, $p_{adj(3)}= .015$, $r=0.475$) and on the 'impossible' trials ($Z=-2.41$, $p_{adj(3)}= .032$, $r=0.407$) compared to the condition in which the change occurred at the actual location of the ball (see **Figure 3.5b**), suggesting that they directed less attention to the empty boxes when the other agent knew where the ball was.

With respect to the UB trials, Friedman test revealed a marginally significant difference between the three types of locations in the number of misses ($\chi^2(2)=5.45$, $p= .065$, Kendall's $W=0.078$), reflecting a pattern similar to the one observed on the TB trials. Yet none of the pairwise comparisons were significant (all Zs<1.40 and all p_{unadj}s> .162).

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

To investigate whether the observed pattern of results was stable over time, we ran a series of Friedman tests, comparing the three conditions separately for first and the second half of the trials and for the two belief types. Regarding the TB trials, although the number of misses was much lower for changes at the actual location of the ball than for changes at either of the two empty locations throughout the task (see Supplementary Materials **Figure S3.11c** and **d)**, with the overall difference between the three types of locations being was significant for both halves (1st half: $\chi^2(2)=6.32$, $p= .042$, Kendall's $W=0.090$; 2nd half: $\chi^2(2)=6.24$, $p= .044$, Kendall's $W=0.089$), follow-up Wilcoxon Signed Rank Test failed to indicate significant difference between the actual and the other two location types in either half of the task (2nd: actual—impossible $Z=-1.99$, $p_{adj(3)}= .120$, $r=0.336$; actual-possible: $Z=-2.06$, $p_{adj(3)}= .120$, $r=0.348$) 1st: actual—impossible $Z=-1.55$, $p_{adj(3)}= .244$, $r=0.262$; actual-possible: $Z=-1.94$, $p_{adj(3)}= .156$, $r=0.328$). As for the UB trials, the three types of locations differed significantly in the first half of the task ($\chi^2(2)=7.91$, $p= .019$, Kendall's $W=0.113$), reflecting a modest but not significant attentional bias towards the actual and the possible locations (actual-impossible: $Z=-1.54$, $p_{adj(3)}= .264$,

r=0.260; actual-possible: Z=-1.71, $p_{adj(3)}$= .264, r=0.289). There was no such bias present in the number of misses in the second half of the task ($\chi^2(2)$=1.81, p= .406, Kendall's W=0.026).

3.4.2.3 Catch (explicit) trials

With respect to the catch ('explicit') trials, participants had on average 11% invalid TB trials, i.e when they pressed two instead of one button (M_{valid}=0.89, SD_{valid}=0.08) and 10% invalid UB trials (M_{valid}=0.90, SD_{valid}=0.13), i.e. when they indicated one instead of two locations, ignoring the fact that the agent represented two alternatives. The hit rate was at ceiling for both types of trials (>0.99). Importantly, on the majority (73.30%) of the UB trials participants indicated first the actual location of the ball, providing of further evidence for 'reality bias' observed on the experimental trials.

3.4.2.4 Comparison of Experiment 2 and Experiment 1b

Reaction times

To investigate whether Experiment 2 differed from Experiment 1b in how fast participants detected changes at the three types of locations, in the two belief conditions, we performed a 2x2x3 mixed ANOVA, with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject and experiment as between-subject factor. The analysis yielded a significant main effect of location type (F(2, 132)=25.16, p< .001, η_p^2= .282), but no significant main effect of belief (F(1, 66)=0.32, p= .576, η_p^2= .005) or significant belief x location type interaction (F(2, 132)=1.01, p= .366, η_p^2= .015). Importantly, despite the markedly different reaction time patterns in the two experiments and the fact that participants were substantially faster to detect changes in Experiment 2 than in Experiment 1b, in general, there was neither a significant main effect of experiment (F(1, 66)=0.38, p= .542, η_p^2= .006) nor a significant experiment x location type (F(2, 132)=0.29, p= .750, η_p^2= .004), experiment x belief (F(1, 66)=1.29, p= .261, η_p^2= .019) or experiment x belief x location type interaction (F(2, 132)=0.61, p= .543, η_p^2= .009). There was no significant difference in the difference scores either (TB diff score: t(66)=0.25, p_{unadj}= .803, d=0.059; UB diff score: t(66)=-1.15, p_{unadj}= .255, d=0.23. As can be seen on **Figure 3.7c** in Experiment 2 the heterogeneity, was similar to the one observed in the previous experiment, both in terms of the presence and the magnitude of the predicted effect.

Miss ratios

Despite the generally lower number of misses in Experiment 2 compared to Experiment 1b, Mann-Whitney tests revealed no significant difference between the two experiments in either of the six main experimental conditions (all *U*s>450.00, all p_{adj}s> .231).

3.4.3 Discussion

The results of Experiment 2 indicate that our paradigm can indeed capture the effect of another agent's underspecified belief content on adults' own representation of the current state of the world: participants, who had to track the other agent's belief deliberately, to be able to respond based on it when asked to do so, were significantly faster in detecting changes not only at the actual but also at the 'possible' location of the ball (compared to the impossible locations) when the other agent did not know in which of the two boxes the ball has hidden, but not when the other agent witnessed the hiding. The effect was present throughout the task and, in some respect, was also reflected in the number of misses, in form of a less pronounced attentional bias towards the actual location of the ball in the UB condition. Such findings imply that, once computed, the underspecified content of another agent's belief influences human adults' behaviour, even when participants have to provide a response where this content is completely irrelevant, generating an effect similar to the one observed in case of false beliefs, i.e. well-specified beliefs and, according to more recent results, even in case of unspecified belief contents (see: Hegedűs &Király, 2022; Kovács et al., 2010; Schneider, Bayliss, et al., 2012).

Importantly, although the predicted difference emerged in Experiment 2, and it was evident that participants continuously monitored the other's belief, according to the instructions, based on their high performance on the incidental catch (i.e. 'explicit belief') trials, the effect, more specifically, evidence for representing the two alternatives from third- and not just from first-person perspective, (quantified as: faster reaction to changes to the possible compared to the impossible location on the UB but not on the TB trials), was present in only about half of the sample and was clearly absent in one-third of the participants. It is most likely that this heterogeneity is the reason why Experiment 2 did not differ significantly from Experiment 1b, despite the apparent difference in the reaction time patterns. These results suggest that the effect of such a belief content on adults' own representation may be not strong enough to emerge in everyone at the timepoint of the measurement, either because (a) this content gets inhibited, to be able to focus on the actual task; (b) the representation of the

other's alternatives fades away; or (c) because belief contents that do not make future events (i.e. the other's behaviour) clearly predictable do not have a reliable effect on spatial attention. This, in turn, raises the question: what do the findings of Experiment 1b, where participants were not instructed to track the other agent's beliefs and we did not find the predicted effect, actually reflect? They may demonstrate that human adults, in general, do not compute other agents' underspecified belief content spontaneously. However, they may also reflect methodological issues with the paradigm that make it difficult to capture the possibly weaker effect of the other agent's belief content on participants' performance - either the presence of possible confounds (see below) or general issues with the sensitivity of the measure.

One potential reason for not being able to reliably demonstrate the effect of the other's belief content on participants' performance, in our implicit experiments, is that we try to measure its impact long after the content has been computed. Previous findings, from a study using an incidental false-belief task, suggest that the time elapsed between the event that presumably triggered the belief attribution and the measurement may influence whether the represented content can exert its impact on the speed of participants' responses. Specifically, if this delay is too long, the content may not be sustained in working memory, hence, its effect may not emerge in participants' responses[16] (Cohen & German, 2009; for further discussion see: Carruthers, 2017). If we assume that participants in our task compute the content of the other's belief and perform the belief attribution on the UB trials in a prospective manner, either when the avatar turns away or when reality changes and, consequently, a difference emerges between the self- and the other perspective, then, on the 'underspecified belief' trials, 8-11 seconds elapse between the computation of the agent's belief content and the timepoint of measurement. During this period the content may easily fade away from working memory, which, in turn, may result in no effect, even if the content was originally computed.

Another possibility is that participants' own knowledge generated such a strong attentional bias towards the actual location of the ball that used up all their available attentional resources. Hence, they could not allocate more attention to the possible location, despite encoding the other agent's underspecified belief. Indeed, the 'actual location bias' was rather strong even in this experiment when the other's belief content was voluntarily tracked. The phenomenon called 'pull-of-the-real' or 'reality bias' is well known in the ToM literature (Carpenter et al., 2002) and it is clearly present not only in children, under age four (see e.g. Birch & Bloom, 2003), but also in adults, as indicated by their looking behaviour in anticipatory looking tasks (see e.g. Schneider et al., 2014; Wang & Leslie, 2016). In fact, in many ToM studies that measure anticipatory looking to the belief-congruent and incongruent

[16] Although it is not entirely clear, from these results, whether it is the time itself or the number of intervening events what matters.

location, the object is removed from the scene, specifically to eliminate this bias, i.e. decrease the executive demands of the task and thereby facilitate the attribution of a reality-incongruent belief to the other (see e.g. Thoermer et al., 2012; Wang & Leslie, 2016)[17]. Although the validity of anticipatory looking as a measure of false belief understanding has been seriously questioned lately (see e.g. Baillargeon et al., 2018; Kampis et al, 2021), a number of other findings (mainly from explicit ToM tasks, see: Carpenter et al., 2002; Setoh et al., 2016), indicate that the removal of the object may help children taking the other's perspective, which, in turn, suggests that it may also affect where adults' attention is directed, in situations where the other agent holds a (partially) discrepant belief.

Finally, it may happen that the stimuli used in Experiment 1b, in their current form, may have prompted some participants to believe that the other agent may know where the ball is, instead of inducing the belief that the agent represents two alternatives when he did not witness the hiding event. In particular, the fact that the agent turned away when the ball already started to move towards its final location, the purpose of which was to trigger the attribution of a belief about the ball's (future) location, might have resulted in the ascription of a true belief, in some participants. Given that the ball's final location was indeed predictable from the moment the belief attribution phase started and humans' well-known tendency to attribute their own knowledge about the outcome of events to others (see e.g.: Birch & Bloom, 2007), as well as informal verbal feedbacks provided by a couple of participants at the end of the experiment, we cannot fully exclude that this feature of the stimuli played some role in the null results we observed in Experiment 1b.

To investigate some of the above-listed possibilities, as a next step, we ran two follow-up experiments, one to test the effect of the timing (Experiment 3), another one to examine the potential effect of the ball's presence and the predictability of its final location (Experiment 4). In Experiment 3, ran to investigate whether reducing the time interval between the potential encoding of the other's belief content and the measurement has an impact on whether or not the predicted attentional bias emerges, we switched the order of the 'return' and the 'outcome' phase in our stimuli, both on the true and the underspecified belief trials, to shift our measurement to an earlier timepoint. Specifically, we altered the trial structure such that, on the underspecified belief trials, the outcome phase started right after the events potentially triggering the belief attribution have ended.

In Experiment 4, to investigate the role of reality bias and the predictability of the ball's final location, we implemented more radical changes in the design: on half of the trials the ball left the scene (to remove the pull-of-the-real), while on the other half it hid in one of the large boxes, with the agent either witnessing this event (true belief) or not (underspecified belief trials), as before. Importantly,

[17] It is worth noting that this way, in these studies participants actually see two false belief scenarios, as the agent mistakenly believes the object to be present even if she witnesses the location change.

we also altered the movement of the ball in the initial phase of the videos as well as the timing of the agent's turn such that on the underspecified belief trials neither the participants nor the agent could infer where the ball would hide (or whether it would exit the scene) at the timepoint when the agent started to turn away from the scene, thus when belief attribution could first take place.

3.5. Experiment 3

Experiment 3 was a modified version of Experiment 1b, differing in the timing of the events: on both the true and the underspecified trials the dot change (or the sound to which participants had to respond on the catch trials) occurred before the agent turned back. In particular, on the underspecified belief trials it occurred after the ball finished jumping and the hiding box's lid closed, that is at the earliest timepoint it was possible to measure the effect of the computed belief content on participants' responses. Note that, on the true belief trials, due to the different timing of the agent's turn (necessary to be able to ensure that the avatar has the same knowledge as the observer), the change/sound occurred later, after the agent has finished turning away (upon hearing the telephone ring). That is, in this version of the study, the two types of belief trials were not matched with respect to the timing of the measurement, i.e. when exactly participants had to provide their responses. They were matched, however, in a sense, that on both trial types, the measurement took place two seconds after the last event that could capture participants' attention before the agent's reappearance. Note that for our hypothesis the within-belief comparisons are crucial and the videos belonging to the three types of locations within the same belief condition were matched in all respects, including the timing of events.

3.5.1 Methods

3.5.1.1 Participants

Participants were 36 university students (*Mage*=21.89, $SD_{age=}$2.36, 13 males), recruited via a student job agency and the university's research participation system (the SONA systems). All participants were right-handed and had normal or corrected-to-normal vision as well as normal hearing. The target sample size was determined to match that of Experiment 1b, even after a potentially high (20%) exclusion rate (expected because of a change in the circumstances of testing, see below). Additional 5 students were tested, but their data was not analysed, either due to a technical error ensuing during the test phase, (N=2) or because participants did not meet the inclusion criteria for the study (miss ratio was >0.30 in one of the conditions: N=1; failed to perform the 'attention check' on >30% of the trials in one of the conditions: N=1-1; did not provide response on the catch trials: N=1). The study was approved by the EPKEB United Ethical Committee, Hungary; participants signed informed consent prior to the experiment and received monetary compensation or gift vouchers for their participation (equivalent to approximately 5 Euros).

3.5.1.2 Stimuli

To bring the measurement closer to the timepoint when the content of the agent's belief was likely computed by the participants, the order of the 'outcome' and the 'return' phase was switched in the test videos. As a consequence, test videos were slightly shorter in Experiment 3 than the ones used in the previous three experiments: they lasted for 22.17 sec, compared to 26.63 sec in Experiments 1a, 1b and 2. Varying the arrangement of the boxes (four), the ball's hiding location (two per arrangement), the direction to which the ball started to move first in the hesitation phase (left/right) and the location of change (four) resulted in 64 different 'change' videos per belief. There were 16 'no-change' videos per belief, one for each scene arrangement (four), hiding location (two) and hesitation direction (two) combination.

Each 'change' and 'no-change' test video had four phases, of which the first and the last were physically identical between the two belief conditions. Just like before, test videos started with a 7.33 sec hesitation phase, which was followed by a 6 sec 'belief attribution' phase, during which the ball jumped into one of the large boxes, with the human avatar either witnessing this event (TB trials) or not (UB

trials). Until this timepoint, specifically, until the box's lid closed at 13.33 sec, the sequence of events was identical to the one participants saw in the previous experiments, with the critical events (the ball's first move, the time when it reached the hiding box and became invisible after the jump) tightly matched between the two types of belief trials. On the UB trials, this event was immediately followed by the 4 seconds long 'outcome' phase. As previously, the outcome phase started with presenting the last frame of the video for 2 sec, after which one of the discs turned red for 750 ms. Following this, the disc turned back to grey again and after 1.25 sec the response period ended. On the TB trials, the 'outcome phase' started after the agent finished turning away at 16.08 sec. Both TB and UB videos ended with the avatar slowly turning back, from the 20.08 sec ('return' phase), i.e., after a 2.75 delay on the underspecified and immediately after the end of the response period on the true belief trials (for details of the timing see **Figure 3.6**).

3.5.1.3 Apparatus, procedure and data analysis

The apparatus was the same as in the previous experiments. The procedure differed from the one used in Experiment 1b in one crucial aspect. As a result of the change in the stimuli, participants had to provide responses 6.3 sec earlier during the UB and 3.55 sec earlier during TB test videos than before. In specific, the onset of the change/sound (which marked the onset of the response period) was 15.33 sec on the UB and 18.08 sec on the TB trials. Consequently, button presses did not end or speed up the trials – the human avatar turned back, and the videos ended at exactly the same timepoint, independent of whether and when participants provided a response. Note that, due to the relocation of the labs, approximately 40% of the participants (N=16) were tested in a different room. All other aspects of the procedure (the instruction, the counterbalancing, and the rules of pseudorandomization) were the same as in Experiment 1b. The data was analysed as before, using the same exclusion criteria.

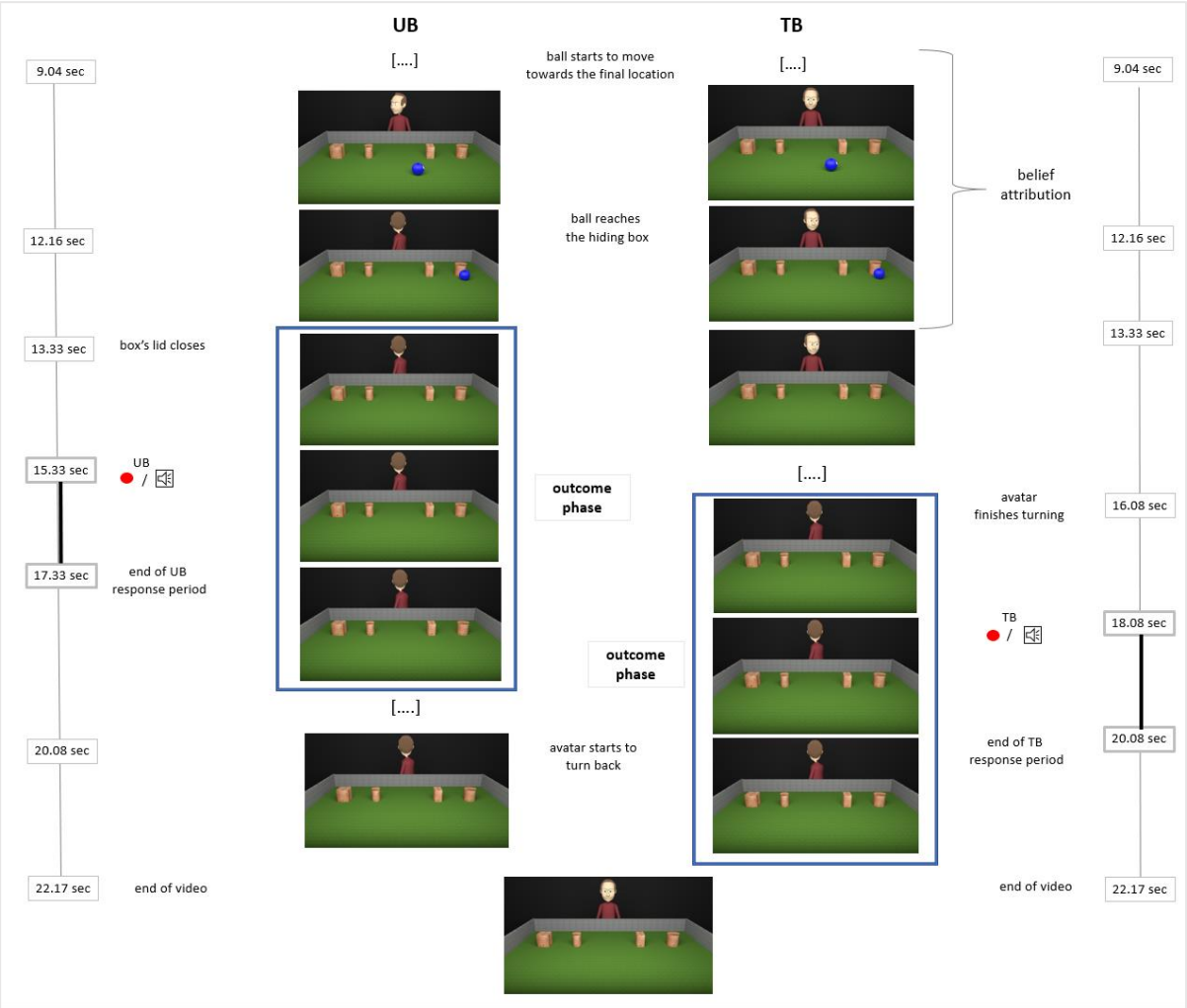


Figure 3.6. Trial structure in the belief attribution, outcome and return phase of the test videos in Experiment 3, on the true (TB) and underspecified belief (UB) trials (the hesitation phase which preceded the belief attribution phase, is not presented on the figure). The two types of belief trials differed (1) in the timing of the avatar's turn, during the belief attribution phase, which defined which events he witnessed and (2) in the timing of the 'outcome phase', that is when the change/sound, to which participants had to react, occurred. On the UB trials, the outcome phase started right after the events presumably triggering the belief attribution ended. On the TB trials, it started after the agent finished turning away. The response period (indicated with a thick black line) started with the onset of the change/sound and lasted for 2000 ms. Responses did not end the test videos, at the end of which the avatar always turned back.

3.5.2 Results

3.5.2.1 Reaction time analyses

Main analyses

There was no significant difference between how fast participants reacted to changes at the two types of 'impossible' locations (TB: $t(35)=-0.82$, $p_{unadj}= .420$, $d=0.136$; UB: $t(32)=1.17$, $p_{unadj}= .249$, $d=0.196$). Therefore, data from these two types of trials were collapsed, and all subsequent analyses were run with three location types (actual, possible, impossible).

As can be seen on **Figure 3.7**a, RTs were relatively high, in all conditions, compared to the ones observed in the previous experiments. A 2 x 3 repeated-measures ANOVA, with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors[18], yielded a significant main effect of location type ($F(2, 70)=17.96$, $p< .001$, $\eta_p^2= .339$), and a tendency level main effect of belief ($F(1, 35)=3.33$, $p= .077$, $\eta_p^2= .087$), reflecting that participants were faster to detect changes at the actual location of the ball than at either of the two other empty locations, independent of belief, and somewhat faster, in general, on the underspecified compared to the true belief trials. Despite this latter difference was present only for changes at the actual and the impossible locations, the belief x location type interaction was not significant ($F(2, 70)=0.13$, $p= .879$, $\eta_p^2= .004$). Crucially, contrary to our expectations, the two belief conditions did not differ in terms of how fast participants detected changes at the possible location. In line with this, pairwise comparisons revealed a significant difference only between the actual and the other two (possible/impossible) locations, on both types of belief trials (TB: actual-possible: $t(35)=-3.39$, $p_{adj(3)}= .004$, $d=0.564$; actual-impossible: $t(35)=-5.36$, $p_{adj(3)}< .001$, $d=0.893$; UB: actual-possible: $t(35)=-2.90$, $p_{adj(3)}= .012$, $d=0.484$; actual-impossible: $t(35)=-3.77$, $p_{adj(3)}= .003$, $d=0.629$), but no significant difference between the possible and the impossible location on the UB ($t(35)=-0.75$, $p_{adj(3)}= .456$, $d=0.126$) and only a tendency level difference on the TB trials ($t(35)=-1.70$, $p_{adj(3)}= .098$, $d=0.284$).

There were 16 (44%) participants whose difference scores (impossible-possible RT) were in line with our predictions, i.e. for whom there was some evidence that they may have represented the two

[18] 2 x 2 x 3 mixed ANOVA, with belief and condition as within-subject factor and testing location as a between-subject factors, revealed no significant effect of testing location (location main effect: $F(1, 34)=1.15$, $p= .290$, $\eta_p^2= .033$; location x belief: $F(1, 34)=0.29$, $p= .295$, $\eta_p^2= .008$; location x condition: $F(2, 68)=1.33$, $p= .271$, $\eta_p^2= .038$; location x belief x condition: $F(2, 68)=0.82$, $p= .446$, $\eta_p^2= .023$), therefore the two sets of data were analysed together.

alternatives from third-person perspective (see Supplementary Materials **Figure S3.7d**). Importantly, there were 15 (41.7%) participants who did not show the predicted effect on the UB trials (had longer RTs for changes at the possible compared to the impossible locations). Even in case of those who did, the magnitude of the effect was rather small and varied to a large extent.

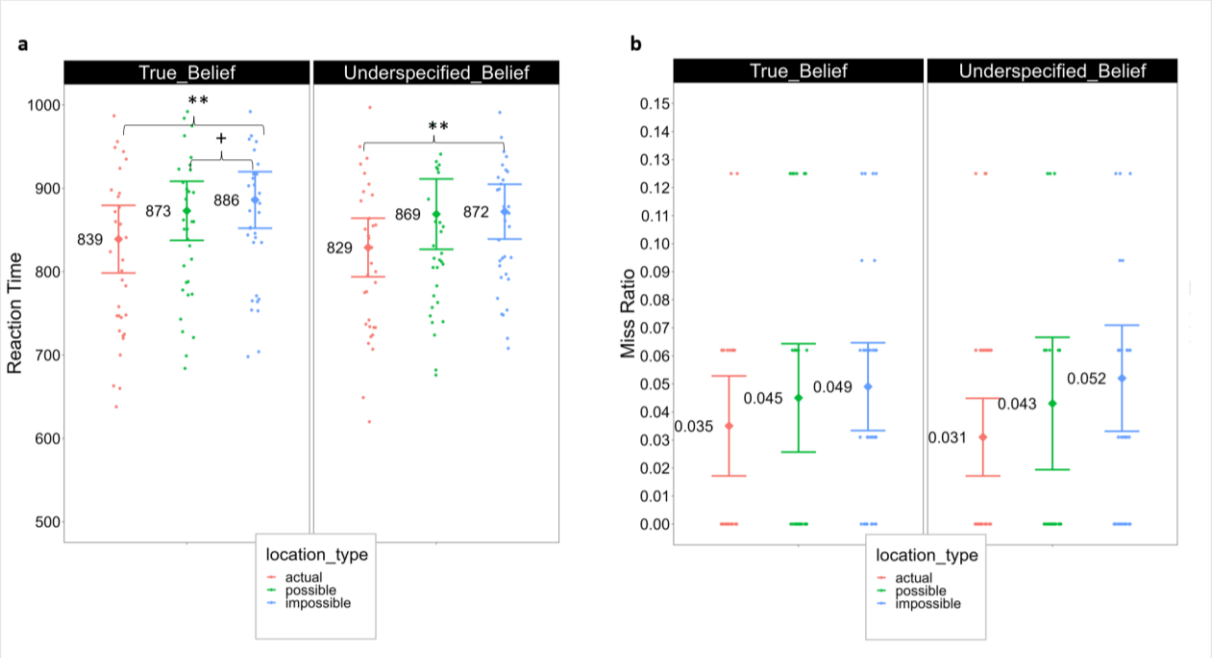


Figure 3.7. (a) Mean reaction time and (b) miss ratio in the true and underspecified belief conditions, for changes occurring at the actual, possible and impossible locations in Experiment 3 (investigating the effect of the timing of the measurement). Error bars represent 95% CI, dots show the individual means. Note: Only the significance levels of the two target comparisons are indicated (actual versus impossible and possible versus impossible) on the figures. For the other comparisons see the main text. The statistical tests for the RT differences were run on the log-transformed RT data. +: p<0.1; *: p<0.05, **: p<0.01

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

A 2 x 2 x 3 repeated measures ANOVA with time (1st versus 2nd half), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors revealed a significant main effect of time ($F(1, 35)=27.71$, $p< .001$, $\eta_p^2= .442$), resulting from the fact that participants were faster in the second than in the first half of the task. There was also a significant main effect of location type ($F(1.76, 61.66)=17.98$, $p< .001$, $\eta_p^2= .339$), reflecting the faster responses for changes at the actual location of the ball (see Supplementary Materials **Figure S3.12a** and **b**), but only a tendency level main effect of belief ($F(1, 35)=2.95$, $p= .095$, $\eta_p^2= .078$) and no significant belief x location type interaction ($F(2,$

70)=0.55, p= .946, η_p^2= .002). Importantly**,** none of the interactions with time were significant (time x belief: F(1, 35)=1.26, p= .269, η_p^2= .035; time x location type: F(1.70, 59.64)=1.67, p= .200, η_p^2= .046; time x belief x location type: F(1.49, 52.22)=1.24, p= .289, η_p^2= .034), indicating that there was no substantial change in the pattern of reaction times throughout the task.

3.5.2.2 Hit rate and miss ratio analyses

Main analyses

The hit rate was at ceiling (all averages>0.99), with no significant difference between the three location types, on either the TB (χ^2(2)=2.80, p= .247, Kendall's W=0.09) or the UB trials (χ^2(2)=2.00, p= .368, Kendall's W=0.028).

Friedman test, performed on the TB trials, indicated no significant difference between the three types of locations in terms of the number of misses (χ^2(2)=4.15, p= .126, Kendall's W=0.058), despite the miss ratio was somewhat lower in the actual than in the other two types of locations (see **Figure 3.7b**). In contrast, there was a significant difference between the three location types on the UB trials, with respect to how many times participants failed to detect changes at those (χ^2(2)=6.49, p= .039, Kendall's W=0.090), with the pattern suggesting that participants directed less attention towards the impossible than to the other two locations, in particular, to the actual location of the ball. Despite the apparent difference between the number of misses, after adjusting for multiple comparisons, the difference between the actual and the impossible (Z=-2.11, $p_{adj(3)}$= .100, r=0.352) and between the actual and the possible condition did not reach significance (actual-possible: Z=-1.36, $p_{adj(3)}$= .176, r=0.226).

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

Regarding the TB trials, although a modest attentional bias towards the actual location was present throughout the task, Friedman tests indicated a marginally significant difference between the three types of locations only on the first (χ^2(2)=5.23, p= .073, Kendall's W=0.073) but not on the second half of the trials (χ^2(2)=1.64, p= .434, Kendall's W=0.023). Despite the three location types clearly differed in terms of the number of misses on these trials, with participants missing the most changes at the possible and the least at the actual location of the ball (see Supplementary Materials **Figure S3.12c**

and **d**), follow-up Wilcoxon Signed Rank Tests failed to indicate significant difference between the actual and the other two types of locations (actual-possible: Z=-1.66, $p_{adj(3)}$= .291, r=0.277; actual-impossible: Z=-1.45, $p_{adj(3)}$= .296, r=0.242). With respect to the UB trials, the pattern was similar to the one observed on the TB trials. There was a tendency level overall difference between the three types of locations ($\chi^2(2)$=4.79, p= .091, Kendall's W=0.067), resulting from the somewhat (though not significantly) lower number of misses for changes at the actual compared to the other two locations in the first half of the task (actual-possible: Z=-1.27, $p_{adj(3)}$= .618, r=0.212; actual-impossible: Z=-0.88, $p_{adj(3)}$= .756, r=0.147), but there was no such difference in the second half of the task ($\chi^2(2)$=3.39, p= .183, Kendall's W=0.047), despite the presence of a modest actual/reality bias.

3.5.2.3 Catch trials

As for the catch trials, the hit rate was at ceiling (averages>0.99 for both the TB and the UB trials). The number of misses was also very low (TB: M=0.010, SD=0.028; UB: M=0.004, SD=0.015), comparable to the number of misses observed in Experiment 1b.

3.5.2.4 Comparison of Experiment 3 and Experiment 1b

Reaction times

A 2x2x3 mixed ANOVA, with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject and experiment as between-subject factor, run on the log-transformed RT data, yielded a significant main effect of location type (F(1.80, 120.49)=28.93, p< .001, η_p^2= .302) and belief (F(1,67)=4.93, p= .030, η_p^2= .069) but no significant belief x location type interaction (F(2, 134)=0.16, p= .850, η_p^2= .002). Importantly, there was a significant main effect of experiment (F(1, 67)=4.83, p= .031, η_p^2= .067), resulting from the fact that participants were much slower to detect changes in Experiment 3 than in Experiment 1b, independent of belief and location type. There was no significant experiment x location type (F(2, 134)=0.002, p= .998, η_p^2= .000), experiment x belief (F(1, 67)=0.01, p= .931, η_p^2= .000) or experiment x belief x location type interaction (F(2, 134)=0.12, p= .884, η_p^2= .002), reflecting the similar reaction time pattern in the two experiments. There was no significant difference in the difference scores either (TB diff score: t(67)=0.19, p_{unadj}= .849, d=0.046; UB diff score: t(67)=-0.21, p_{unadj}= .831, d=0.052).

Miss ratios

Despite the somewhat lower miss ratios, on the TB possible and impossible trials and the somewhat higher number of misses on the UB possible and impossible trials in Experiment 3 (compared to Experiment 1b) Mann-Whitney tests revealed no significant difference between the two experiments in either of the six main experimental conditions (all *U*s>525.00, all $p_{adj(3)}$s> .245).

3.5.3 Discussion

Despite performing the measurement as close to the timepoint of the potential belief attribution as possible, Experiment 3 found no evidence for the spontaneous tracking of the other agent's hypotheses. The reaction time pattern was similar to the one observed in Experiment 1b: participants were faster to detect changes at the actual location of the ball, independent of the agent's belief, but, contrary to our expectations, they did not react faster to changes at the possible but empty location (and also did not miss fewer changes there), when the avatar was uncertain about where the ball has hidden. Importantly, this does not mean that the results were the same as the ones obtained in Experiment 1b. In fact, they differed from those in multiple aspects. First, unlike in Experiment 1b, where participants missed roughly equal number of changes at the three locations when the agent was uncertain, in this experiment 'reality bias' emerged not only in the true but also in the underspecified belief condition, in terms of the number of misses. Second, participants' reaction times were longer in this than in any of the previous experiments, independent of the location of change and the belief of the agent.

Although the most likely interpretation of our findings, specifically the lack of attentional bias towards the possible location of the ball (on trials where the agent represented two alternatives), is that, in the absence of any external prompt, participants simply did not represent the other agent's hypotheses, the observed differences between the two experiments raise the possibility that our manipulation might have had the opposite of the intended effect. Specifically, it might have made it more (and not less) difficult to capture the effect of any attributed content on the attention of participants. For instance, the fact that participants had to respond to the change soon after the ball finished hiding, on the underspecified belief trials, might not have left enough time for them to disengage from the box

where the ball has hidden, preventing them from monitoring the other, 'possible' location. This may also explain the presence of the reality bias on these trials, in the number of misses.

Bringing the measurement to an earlier timepoint might have left much less time for participants to 'prepare' for their response, in general, after the last event to which they had to react, i.e. the agent's turn, has ended. This may have made the task generally more difficult, which, in turn, may explain why participants were slower in Experiment 3 than in any of the previous experiments. Investing excessive attentional resources in the actual task may have masked the effect in itself, generating a situation in which, even if participants represented the content of the other agent's belief in some form, performing the attribution before the other would have turned away, this representation could not exert its effect on how they allocated their attention in the experiment. Finally, since in this version of the task buttonpresses did not end the trials the way they did before, to keep the length of trials constant, participants may have also been less motivated to act as fast as possible than those in the previous experiments. Slower responses, in turn, may have masked the modulatory effect of the other's belief content, even if it did affect participants' attention to some extent.

In any case, the results of Experiment 3 indicate that the null results in Experiment 1b were most likely not due to the timing of the measurement. It reflects either the fact that people do not track the hypotheses of other agents spontaneously, when those are not relevant for them in some sense, or the limitations of the paradigm to capture the potentially very small effect of another agent's spontaneously computed underspecified belief content on participants' own representation of the state of affairs, due to other factors playing an important role in where spatial attention is directed, such as the reality bias and/or the predictability of the ball's final location, or general issues with the sensitivity of our measure.

3.6. Experiment 4

Experiment 4 tested two hypotheses regarding the potential causes of the previous null results observed in Experiment 1b: that the potential effect of belief tracking on participants' attention was masked by (1) the pull-of-the-real and/or (2) characteristics of the stimuli, that may have made the outcome predictable. Consequently, it differed from Experiment 1b in two main respects. Below, we summarize these differences.

First, to eliminate the 'reality bias', on half of the trials the ball left the scene after the hesitation phase ('ball absent' trials), while on the other half, it hid as before ('ball present' trials), with the agent either

watching the hiding/leaving event (true belief trials) or not (underspecified belief trials). The main reason for including the latter condition, i.e. the ball present trials, was to ensure that participants would have reason to assume that the other agent represents the two large boxes as potential hiding places for the object on those trials where he lacks certain knowledge about the outcome of the previously observed events. We could have achieved this simply by removing the ball after the agent has turned away, such that only the participant knows that the ball is not present. However, this would have rendered the agent's belief false on all trials (independent of what events he had witnessed). We wanted to avoid this to be able to compare our results to those of the previous experiments on the one hand and to have conditions, for which we have clear predictions, on the other. Importantly, such a design meant that on the underspecified belief trials the agent actually represented three alternatives, making the to-be-attributed content somewhat more complex than in the previous experiments ('*the ball is either at location A or location B or it left the scene'*).

Second, we also changed the movement of the ball in the first half of the belief induction phase as well as the timing of the avatar's turn, to make the ball's final location less predictable, both from first- and third-person perspective. In Experiment 1a, 1b and Experiment 2, the fact that the ball has already started to move towards the hiding box when the agent started to turn, may have induced the sensation in participants that the other already collected sufficient information to predict where the ball will hide, thus, he may infer its final location even in the absence of direct visual evidence. At least, this may have made it easier for participants to (mistakenly) attribute the other agent what they could infer from the ball's movement at this very moment. To minimize the possibility that participants attribute a true belief to the agent even when he did not witness the hiding, we altered the stimuli such that when the agent started to turn away on the underspecified trials the outcome (whether the ball will hide or not and if yes, in which of the two boxes) was still completely unpredictable both for the participants and the other agent.

We hypothesized that if adults spontaneously represent the alternatives represented by another agent and if the lack of a difference between the possible and the impossible condition in Experiment 1b was indeed due to a strong reality bias, the expected effect should emerge in the 'ball absent' but not in the 'ball present' condition. That is, RTs should be significantly shorter in this case for changes at the possible than for changes at the impossible location(s), on trials where the avatar lacks knowledge about the outcome, with no difference between the speed of responses for changes at the two 'possible' locations. If previous results were rather due to the predictability of the ball's final location, the expected difference should emerge on the 'ball present' trials, and, depending on whether or not the other's hypotheses are represented when the object is not present in the scene, it should either be present or absent on the 'ball absent' trials.

3.6.1 Methods

3.6.1.1 Participants

Participants were 35 university students (M_{age}=22.37, SD_{age}=2.91, 16 males), recruited via a student job agency and the university's research participation system (the SONA systems). All participants were right-handed and had normal or corrected-to-normal vision as well as normal hearing. Additional two students were tested but their data was excluded because the participants did not meet the inclusion criteria for the study (miss ratio was >0.30 in at least one of the six main conditions of the 'ball present' or one of the four main conditions of the 'ball absent' trials). The target sample size was determined to match that of Experiment 1b.

The study was approved by the EPKEB United Ethical Committee, Hungary; participants signed informed consent prior to the experiment and were compensated as in Experiment 3.

3.6.1.2 Stimuli

Test videos consisted of Belief induction and Outcome videos, as in the earlier studies. Outcome videos were identical to those used in Experiment 1a, 1b and 2. Belief induction videos were 19.96 sec long and differed from those used in the first three experiments in the hesitation and belief attribution phase (see below). Depending on the event type (i.e. the outcome of the event sequence) belief induction videos could either be 'ball present' (the ball hid in a box) or 'ball absent' videos (the ball left the scene). Half of the 'ball present' and the 'ball absent videos' depicted a 'true' belief scenario (TB videos), half of them an 'underspecified' belief scenario (UB videos), which varied in which events the human avatar witnessed (as before). Both TB and UB videos had eight versions per event type, varying in the scene arrangement, and the direction (box) towards which the ball started to move first (left/right, as well as in its final location, on the 'ball present trials').

Each video, independent of the event type and the agent's belief, had three phases with the first and the last phase being physically identical between the two belief conditions, as before. They started with a 6.46 sec long 'hesitation' phase, in which the ball moved first in one direction, then turned back and started to move towards the middle. This phase was followed by a 11.21 sec long 'belief attribution' phase, during which the ball could continue its movement in three ways: after reaching the middle of the screen it could (1) either move on to hide in the box towards which it headed at the start of the belief attribution phase; (2) turn back to hide in the first box it approached ('ball present',

continued move vs turnback trials, 1 and 2, respectively); or (3) first continue its move then turn back and finally exit the scene in the middle lower part of the screen ('ball absent' trials), while the human avatar was either facing the scene or not. Videos ended with the avatar turning back, from the 17.87 sec ('return' phase). The two belief conditions differed only in the timing of the avatar's turn. In the UB videos, it started right after the end of the hesitation phase and almost finished by the time the ball reached the middle of the screen (at 7.67 sec). In the TB condition, the avatar started to turn after the ball finished jumping in the box and became invisible, as before, at 13.59 sec. Apart from this difference, the two belief conditions were tightly matched with respect to the timing of the critical events, just like in the previous experiments (for details see **Figure 3.8**). Belief induction videos were immediately followed by one of the outcome videos: a 'change outcome' video on the experimental and a 'no-change outcome' video on the 'catch' trials. Importantly, 'no-change outcome' videos were only paired with 'ball present' trials, as the question about the ball's location made sense only in this case but not when it left the scene.

3.6.1.3 Procedure

The trial structure was the same as in Experiment 1b as well as the participants' task and the instruction, with the only difference being that they were told that they would have to indicate the location of the ball only when the ball hides in one of the boxes (i.e. on the 'ball present' trials).
Depending on the hiding location of the ball and the location of the change, experimental trials could belong either to the actual, possible, impossible_R (rectangle shape) or to the impossible_C (cylinder shape) experimental condition on the 'ball present' or to the possible1, possible2, impossible_R or to the impossible_C experimental condition on the 'ball absent' trials, where possible1 refers to the location that was first approached by the ball, possible2 to the location towards which it headed second, and R and C denotes the shape of the two small boxes (the cylinder and the rectangle, respectively). Participants received 160 test trials in four blocks: 16 'ball present' 16 'ball absent' experimental and 8 'ball present' catch trials per block, which were randomly intermixed with the experimental trials. Each block contained two experimental trials per event type, belief and location type (specifically, one 'continued move' and one 'turnback' trial per belief and location type on the 'ball present' trials), and four catch trials per belief type. The arrangement of the boxes, the hiding location of the ball and the location of change were counterbalanced within event type, belief and block. The order of the trials within the blocks was pseudorandomized, such that there were no more than three consecutive trials with the same event type, belief, location type, arrangement, and no more than two consecutive trials with the same hiding or change location.

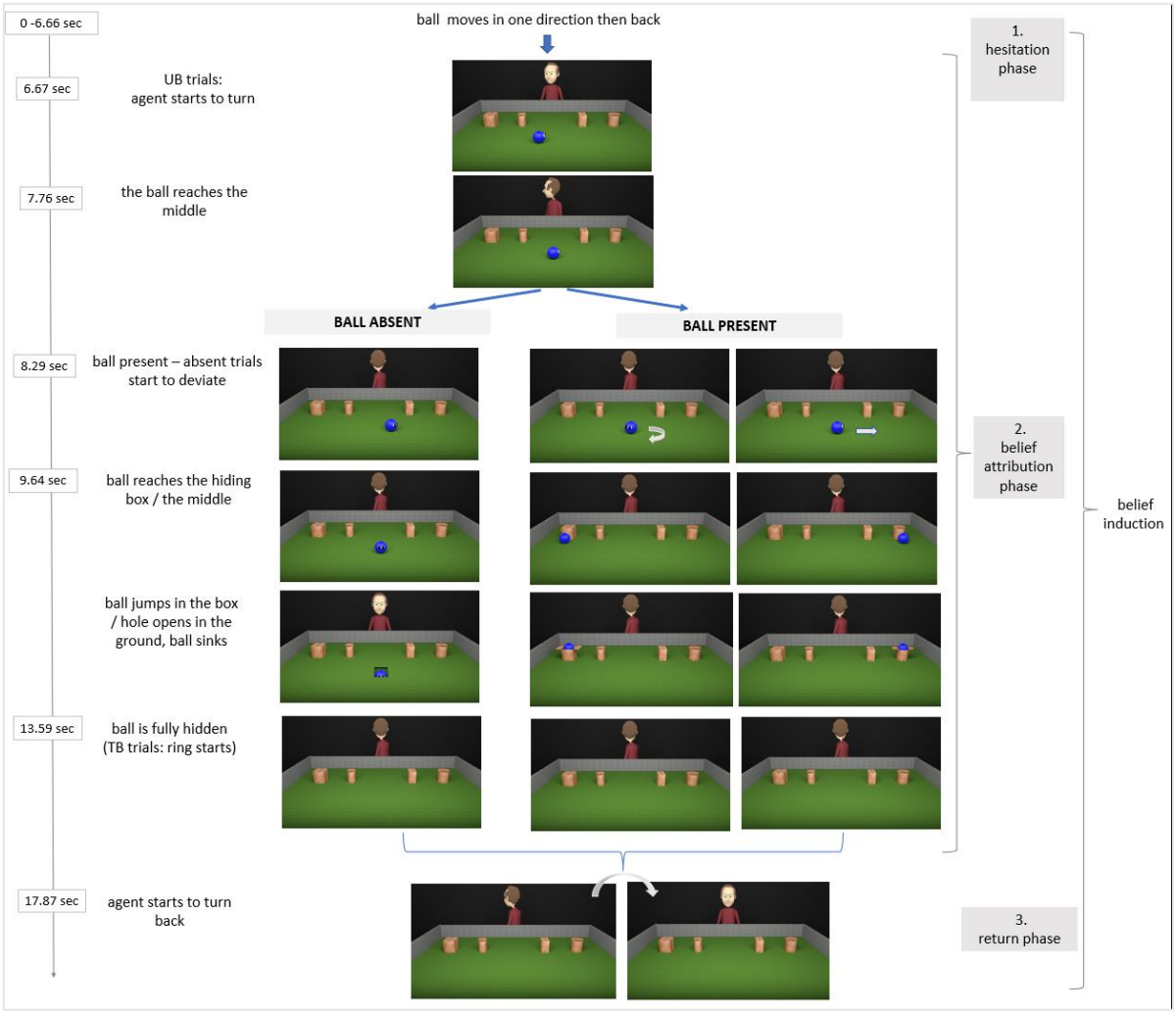


Figure 3.8. Trial structure in the belief induction videos in Experiment 4, on the 'ball present' and 'ball absent' underspecified belief (UB) trials. The outcome of the event sequence (hiding in one or the other large box, leaving the scene by sinking in a hole that opened in the middle) was unpredictable at the timepoint when the agent started to turn away. True belief trials differed from the underspecified belief trials only in the timing of the avatar's turn, which took place immediately after the ball became invisible. Belief induction videos were followed by 4 seconds long 'change' or 'no-change' outcome videos, as in Experiment 1a, 1b and Experiment 2. 'No-change' videos were paired only with 'ball present' trials (as the ball did not have a specific location on the 'ball absent' trials).

3.6.1.4 Data analysis

The primary dependent measures and the data exclusion criteria were the same as before (see section 2.1.4). Mean reaction times and miss ratios were calculated per belief and location type, separately for the 'ball present' and the 'ball absent' trials on the log-transformed RT data. Just like in the previous experiments, first, we compared the two 'impossible' locations along the mean RTs (performing the comparisons separately for the 'ball present' and 'ball absent' true and underspecified belief trials), then (as there was no difference) we collapsed the trials over these two types of locations. After a similar comparison, we have also analysed the two possible locations together, on the 'ball absent' trials. This resulted in three main location types per belief on the 'ball present' (actual, possible, impossible) and two main types of locations or experimental conditions per belief on the 'ball absent' trials (possible versus impossible). Participant exclusion criteria were the same as in Experiment 1a and 1b, taking into account not only the miss ratios and the attention check performance on the six main location types of the 'ball present' but also the performance on the four main types of locations of the 'ball absent' trials. The data was analysed as before with all analyses run separately for the 'ball present' and 'ball absent' trials. Our crucial tests were: (a) the comparison of RTs for changes at the actual and the impossible locations in the ball present true belief trials (baseline effect) and (b) the comparison of RTs (as well as the number of misses) for changes at the possible and impossible locations, on the underspecified and the true belief trials, in the 'ball present' and (c) 'ball absent' conditions, with (b) and (c) testing for our two hypotheses. In addition, we also compared the 'ball present' and 'ball absent' trials along (1) the possible and impossible trials' mean RTs and (2) the difference scores. Comparisons to Experiment 1b were run including only the 'ball present' trials.

3.6.2 Results: 'ball present' trials

3.6*.2.1 Reaction time analyses*

Main analyses

Given that there was no significant difference between how fast participants reacted to changes at the two types of 'impossible' locations (TB: $t(34)$=-0.86, p_{unadj}= .395, d=0.015; UB: $t(34)$=1.06, p_{unadj}= .296, d=0.180), data from these two types of trials were collapsed and all subsequent analyses were run with three location types.

A 2 x 3 repeated-measures ANOVA with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors, yielded a significant main effect of location type ($F(2, 68)=25.38$, $p< .001$, $\eta_p^2= .427$) and belief ($F(1, 34)=4.64$, $p= .038$, $\eta_p^2= .120$) but no significant belief x location type interaction ($F(1.54, 52.37)=0.36$, $p= .933$, $\eta_p^2= .001$). As can be seen on **Figure 3.9a**, participants were faster to detect changes not only at the actual but also at the possible location of the ball compared to the impossible locations, independent of belief. In line with this, pairwise comparisons revealed a significant difference between the actual and the other two as well as between the possible and the impossible location on both the TB (actual-impossible: $t(34)=-4.41$, $p_{adj(3)}< .001$, $d=0.745$; actual-possible: $t(34)=-2.48$, $p_{adj(3)}= .034$, $d=0.419$; possible-impossible: $t(34)=-2.52$, $p_{adj(3)}= .034$, $d=0.420$) and the UB trials (actual-impossible: $t(34)=-4.92$, $p_{adj(3)}< .001$, $d=0.832$; actual-possible: $t(34)=-2.25$, $p_{adj(3)}= .031$, $d=0.379$; possible-impossible: $t(34)=-2.58$, $p_{adj(3)}= .030$, $d=0.435$). Importantly, the difference between the possible and the impossible condition's RT was similar for the two beliefs ($t(34)=-0.58$, $p= .565$, $d=0.098$), i.e. there was no difference in the magnitude of the 'effect'. Despite participants were generally faster on the underspecified than on the true belief trials, pairwise comparisons indicated no significant difference between the two types of beliefs at any of the three types of locations either (all ts < 1.72, all p_{adj}s > .287).

Inspection of the individual data revealed that there were only 11 (31%) participants whose difference scores (impossible-possible RT) were in line with our predictions, i.e. for whom there was some evidence that they may have represented the two alternatives from third-person perspective, spontaneously (see Supplementary Materials **Figure S3.7e**). Importantly, the number of those participants who did not show the predicted effect on the UB trials (had longer RTs for changes at the possible compared to the impossible locations) was also relatively low (N=9, 26%).

Figure 3.9. (a) Mean reaction time and (b) miss ratio in the true and underspecified belief conditions of the 'ball present' trials, for changes occurring at the actual, possible and impossible locations. Error bars represent 95% CI, dots show the individual means. Note: Only the significance levels of the two target comparisons are indicated (actual versus impossible and possible versus impossible) on the figures. For the other comparisons see the main text. The statistical tests for the RT differences were run on the log-transformed RT data. +: p<0.1; *: p<0.05, **: p<0.01

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

A 2 x 2 x 3 repeated measures ANOVA with time (1st versus 2nd half), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors revealed a significant main effect of location type ($F(1.76, 59.88)=25.20$, $p< .001$, $\eta_p^2= .426$) and a significant main effect of belief ($F(1,$

34)=6.58, p= .015, η_p^2= .162) but no significant a belief x location type interaction (F(1.55, 52.71)=0.13, p= .970, η_p^2= .000). Importantly, despite a marked decrease in the magnitude of the effect on the underspecified belief trials, from the first to the second half of the task (see Supplementary Materials **Figure S3.13a** and **b**), none of the interactions with time were significant (time x belief: F(1, 34)=0.01, p= .935, η_p^2= .000; time x location type: F(2, 68)=0.19, p= .831, η_p^2= .005; time x belief x location type: F(2, 68)=1.11, p= .335, η_p^2= .032) - the general pattern of reaction times was stable over time. Unlike in the previous experiments, there was no significant main effect of time either (F(1, 34)=0.46, p= .502, η_p^2= .013).

3.6.2.2 Hit rate and miss ratio analyses

Main analyses

Hit rate was at ceiling (all averages>0.99), with no significant difference between the three types of locations (TB: χ^2(2)=1.00, p= .607, Kendall's W=0.014; UB: all means = 1.00).

Despite participants missed more changes at the possible than at either of the other two (particularly the actual) locations on the TB trials (see **Figure 3.9b**), Friedman test indicated no significant difference between the three types of locations (χ^2(2)=1.33, p= .513, Kendall's W=0.019). There was no significant difference between the three types of locations on the UB trials either with respect to how many times participants failed to detect changes at those (χ^2(2)=2.07, p= .355, Kendall's W=0.030): the number of misses was generally very low on these trials, independent of the location of change. Crucially, participants missed fewer changes on the UB than on the TB trials, in general. The difference between the two types of trials was significant for changes at the possible (Z=-2.50, $p_{adj(3)}$=.039, r=0.422), but not for changes at the other two locations (actual: Z=-0.28, $p_{adj(3)}$=.782, r=0.047; impossible: Z=-1.55, $p_{adj(3)}$=.240, r=0.262).

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

Although there was a modest actual bias present on the TB trials, in the first, and on the UB trials in the second half of the task (see Supplementary Materials **Figure S3.13c** and **d**), Friedman tests indicated no significant difference between the three types of locations in terms of the number of

misses in either of the two halves of the experiment, on either of the two types of belief trials (1st half – TB: $\chi^2(2)=3.03$, $p= .202$, Kendall's $W=0.043$; UB: $\chi^2(2)=1.47$, $p= .479$, Kendall's $W=0.021$; 2nd half – TB: $\chi^2(2)=0.30$, $p= .859$, Kendall's $W=0.004$; UB: $\chi^2(2)=4.26$, $p= .112$, Kendall's $W=0.061$).

3.6.2.3 Catch trials

With respect to the catch trials, the hit rate was high (average>0.93 on both the TB and UB trials), although lower than in Experiment 1b, and the number of misses was very low (TB: M =0.007, SD=0.025; UB: M =0.002, SD=0.011), indicating that participants could track the location of the ball and recall this information when needed.

3.6.2.4 Comparison of Experiment 1b and Experiment 4 – 'ball present' trials

Reaction times

A 2x2x3 mixed ANOVA, with belief (TB versus UB) and location type (actual, possible, impossible) as within-subject and experiment as between-subject factor, yielded a significant main effect of location type ($F(1.83, 120.85)=35.35$, $p< .001$, $\eta_p^2= .349$) and belief ($F(1, 66)=6.27$, $p= .015$, $\eta_p^2= .087$) but no significant belief x location type interaction ($F(1.75, 115.70)=0.99$, $p= .882$, $\eta_p^2= .001$). Importantly, there was a significant main effect of experiment ($F(1, 66)=5.43$, $p= .023$, $\eta_p^2= .076$): participants were much faster to detect changes in Experiment 4 than in Experiment 1b, independent of belief and location type. Mean RTs were similar to the ones observed in our very first experiment. Despite the markedly different pattern, none of the interactions with experiment were significant (experiment x location type: $F(1.83, 120.85)=1.31$, $p= .271$, $\eta_p^2= .020$; experiment x belief: $F(1, 66)=0.53$, $p= .469$, $\eta_p^2= .008$; experiment x belief x location type interaction: $F(1.75, 115.70)=0.05$, $p= .935$, $\eta_p^2= .001$). The two experiments did not differ along the difference scores either (TB diff score: $t(66)=1.17$, $p_{unadj}= .245$, d=0.285; UB diff score: $t(66)=-0.94$, $p_{unadj}= .353$, d=0.212).

Miss ratios

In line with the reaction time results, participants in Experiment 4 had generally fewer misses, at all types of locations, than those in Experiment 1b. After adjusting for multiple comparisons, the Mann-Whitney tests revealed a significant difference only for the UB possible (Mdn_{Exp1b}=39.82 ms versus

Mdn_{Exp3b}=29.49 ms, U=402.00, $p_{adj(3)}$= .018) but not for the other trials (all Us > 416.00, all $p_{adj(3)}$s > .113).

3.6.3 Results: 'ball absent' trials

3.6.3.1 Reaction time analyses

Main analyses

There was no significant difference between how fast participants reacted to changes at the two types of 'impossible' locations (TB: t(34)=0.64, p_{unadj}= .528, d=0.108; UB: t(34)=0.31, p_{unadj}= .761, d=0.059), therefore data from these two types of trials were collapsed. There was no significant difference between the two possible locations either, on the underspecified belief trials (t(34)=0.51, p_{unadj}= .959, d=0.009), with the two means being almost the same (possible1: M=722.24, SD=102.47; possible2: M=721.92, SD=121.07). Although in the true belief condition participants were significantly slower to detect changes at one possible location compared to the other, the one that was approached second (possible2 > possible1: t(34)=-2.34, p_{unadj}= .025, d=0.396; possible1: M=742.49, SD=114.54; possible2: M=775.23, SD=117.10), since this difference was rather unexpected and hard-to-interpret on the one hand and our hypothesis concerned the difference between the two possible locations on the underspecified belief trials on the other, data from these trials were also collapsed, and all subsequent analyses were run with two conditions (possible versus impossible). Results of the analyses run with the three location types are reported in the Supplementary Materials **S3.3**).

A 2 x 2 repeated-measures ANOVA, with belief (TB versus UB) and location type (possible, impossible) as within-subject factors, yielded a significant main effect of belief (F(1, 34)=13.22, p= .001, η_p^2= .280), and a tendency level main effect of location type (F(2, 68)=2.88, p= .099, η_p^2= .078), but no significant belief x location type interaction (F(2, 68)=0.56, p= .461, η_p^2= .016): participants were faster to detect changes on the underspecified than on the true belief and on the possible than on the impossible trials, in general (see **Figure 3.10a)**. Crucially, however, the planned pairwise comparisons revealed significant difference between the possible and the impossible locations only on the UB (t(34)=-2.09, p= .044, d=0.353) but not on the TB trials (t(34)=-0.53, p= .597, d=0.090), reflecting the fact that, although the reaction time pattern was similar on the two types of belief trials, the attentional bias towards the possible locations was more pronounced when the agent was uncertain about the ball's

actual location than when he had the same knowledge as the participants, in line with how we predicted. There were 15 (43%) participants for whom there was some evidence that they may have represented the two alternatives from third-person perspective, as indicated by their difference scores (difference score >0 on the UB trials and either a difference score ≤ 0 on the TB trials or a positive score smaller than the one obtained for the UB trials; see Supplementary Materials **Figure S3.7f**). There were N=14 (41.7%) participants who did not show the predicted effect on the UB trials (had longer RTs for changes at the possible compared to the impossible locations).

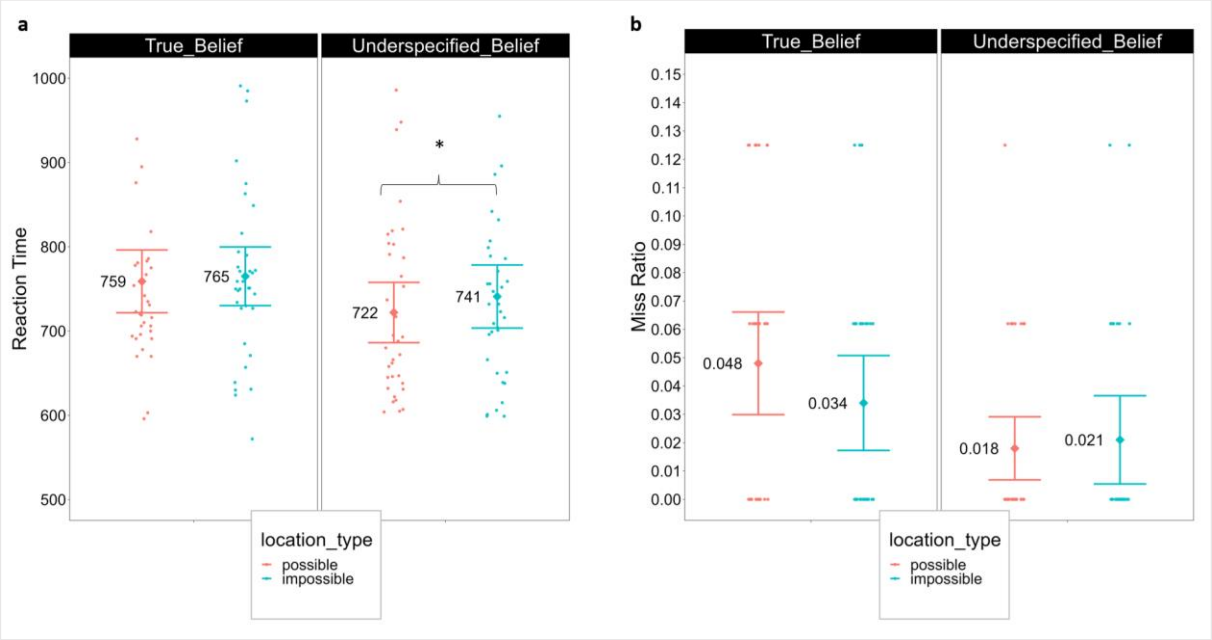


Figure 3.10. (a) Mean reaction time and (b) miss ratio in the true and underspecified belief conditions of the 'ball absent' trials, for changes occurring at the possible and impossible locations. Error bars represent 95% CI, dots show the individual means. The statistical tests for the RT differences were run on the log-transformed RT data. +: p<0.1; *: p<0.05, **: p<0.01

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

A 2 x 2 x 2 repeated measures ANOVA with time (1st versus 2nd half), belief (TB versus UB) and location type (possible, impossible) as within-subject factors yielded a significant main effect of belief (F(1, 34)=11.72, p= .002, η_p^2= .256) and a tendency level main effect of location type (F(1, 34)=3.29, p= .078, η_p^2= .088). There was, however, no significant belief x location type (F(1, 34)=0.50, p= .485, η_p^2= .014), time x belief (F(1, 34)=1.89, p= .179, η_p^2= .053), time x location type (F(1, 34)=0.02, p= .902, η_p^2= .000) or time x belief x location type interaction (F(2, 68)=0.44, p= .510, η_p^2= .013) – the predicted effect was present throughout the task (see Supplementary Materials **Figure S3.14a** and **b**). There was no

significant main effect of time either ($F(1, 34)=0.01$, $p= .929$, $\eta_p^2= .000$), reflecting the fact that, unlike in the previous experiments, participants did not become faster with time.

3.6.3.2 Hit rate and miss ratio analyses

Main analyses

The hit rate was at ceiling (all averages>0.99), with no significant difference between the types of locations (TB: $Z=-1.00$, $p= .317$, $r=0.169$; UB: $Z=-1.41$, $p= .618$, $r=0.238$). In a similar way, Wilcoxon Signed Rank Tests, revealed no significant difference between the possible and the impossible location in terms of the number of misses, on either the true ($Z=-1.06$, $p= .287$, $r=0.179$) or the underspecified belief trials ($Z=-0.30$, $p= .766$, $r=0.051$). Importantly, however, the miss ratio was lower for the UB than for the TB 'possible' trials ($Z=-2.46$, $p_{adj(2)}=.028$, $r=0.416$), while no such difference was present for the 'impossible' trials ($Z=-1.16$, $p_{adj(2)}= .246$, $r=0.196$), indicating that participants allocated more attention to the two boxes the agent could consider a potential hiding place for the ball on those trials when he was uncertain about the ball's actual location (see **Figure 3.10b**).

Additional analyses: comparing the first versus the second half of the task (the effect of 'time')

Wilcoxon Signed Rank Test revealed that in the first half of the experiment, participants missed more changes at the possible than at the impossible locations on the TB trials, though the difference was only marginally significant ($Z=-2.13$, $p_{adj(2)}=.066$, $r=0.360$). No such difference was present in the second half of the experiment ($Z=-0.54$, $p_{adj(2)}=.592$, $r=0.091$) or on the UB trials, in either halves of the task (1st half: $Z=-0.18$, $p_{adj(2)}=.858$, $r=0.030$; 2nd half: $Z=-1.00$, $p_{adj(2)}=.634$, $r=0.169$). Except for the TB impossible trials, participants had generally fewer misses in the second, compared to the first half of the task (see Supplementary Materials **Figure S3.14c** and **d**). Wilcoxon Signed Rank Tests indicated a significant difference only for the UB impossible ($Z=-2.31$, $p_{adj(2)}=.042$, $r=0.212$) but not for the other trials (all Zs <1.52, all ps >.129).

3.6.3.3 Comparison of the ball present and ball absent trials

Comparison of the 'ball present' and 'ball absent' trials along RTs for changes at the possible and the impossible locations, revealed a significant difference between the two types of trials in all but one experimental conditions: participants were significantly faster to detect changes at the impossible locations, on both the TB (t(34)=2.42, $p_{adj(2)}$= .042, d=0.409) and the UB (t(34)=3.65, $p_{adj(2)}$= .002, d=0.617), and at the possible location on the UB trials (t(34)=2.20, $p_{adj(2)}$= .035, d=0.372), when the ball left the scene than when it remained there (though hidden in one of the boxes). Despite the different pattern of results, specifically the presence of the effect in the true belief condition on the 'ball present' but not on the 'ball absent' trials, statistical tests revealed no significant difference between the two types of trials, along either of the two difference scores (TB: t(34)=1.6, p= .183, d=0.230; UB: t(34)=0.5, p= .727, d=0.059).

3.6.4 Discussion

The results of Experiment 4 indicate that, unlike the timing of the measurement (a factor investigated in Experiment 3), the 'pull-of-the-real' might have indeed played a role in the fact that we could not reliably demonstrate an effect of the other agent's belief content on participants' performance in our previous 'implicit' experiments. Eliminating this confound clearly made the task easier for participants, as reflected by their shorter reaction times on the 'ball absent' trials compared to those where the ball remained present. Crucially, it did not only facilitate their attention in general but also led to the emergence of the predicted effect: participants were faster to detect and missed fewer changes at the possible than at the impossible locations on those trials where the other agent was uncertain about the actual state of affair (but not when he had the same knowledge as the participant), indicating that they allocated more attention to these locations when the someone else could consider these a potential hiding place for the object, even though this other agent's belief was irrelevant and the ball was actually absent. These results corroborate previous findings (see e.g. Wang & Leslie, 2016) showing that removing the object from the scene about which the other agent and the participant entertain a different belief can aid the attribution of the appropriate, reality-incongruent belief content to the agent in implicit ToM tasks, presumably by helping disengagement and thereby lowering their executive demands.

It is important to note, however, that, despite the presence of the effect on the 'underspecified belief' trials, the difference between the two crucial conditions was rather small. Relatedly, the individual heterogeneity was very high, both in terms of the presence and the magnitude of the effect.

Consequently, when analysing the two possible locations separately, as in the previous experiments, the difference was not significant. It is even more important that in this case the results of the TB trials were much less clear. In specific, RTs for the two possible locations (the one approached first/last by the ball and the one approached second) seemed to differ markedly, with the pattern indicating attentional bias towards one possible location (towards the box that was approached first by the ball) and/or, possibly, inhibition of the other (see Supplementary Materials **S3.5**). Despite, the separate analysis of the two possible locations meant averaging very few trials per condition, which may mean that both the difference observed on the TB trials (and the lack of difference on the UB trials) might have just been a chance finding, these results warrant caution. Likely more studies are needed to be able to draw firm conclusions regarding when and how people track other agents' hypotheses.

Contrary to our findings on the 'ball absent' trials, in the 'ball present' trials, where the ball did not leave the scene, and the cues which would have made it possible for the agent to predict its hiding location were eliminated, we did not find strong evidence for the tracking of the agent's hypotheses in participants' reaction time pattern, indicating that the predictability of the outcome may not have played a substantial role in our previous null results. Although participants were faster to detect changes at the possible than at the impossible locations, this was true not only on the underspecified but also on the true belief trials, with no difference in the magnitude of the effect. Such a pattern implies that participants represented both alternatives from first- person perspective (at the beginning of the event sequence), then failed to completely inhibit the alternative they could have eliminated, possibly due to a stronger initial encoding of the possibilities, resulting from the higher uncertainty regarding the outcome of the events. Nevertheless, this does not mean that the other agent's belief did not influence participants' attention in any ways when the ball was present. Just like in the 'ball absent' condition, participants were generally faster and missed fewer changes on the underspecified than on the true belief trials, particularly at the location the other could consider 'possible'. This suggests that they may have encoded the avatar's belief content or at least the difference between their own and the other agent's knowledge state, which, in turn, facilitated their change detection performance, even in the face of a strong reality bias.

Altogether results of Experiment 4 indicate that human adults may be sensitive to other agents' beliefs not only when those have a specified but also when they have an underspecified belief content, or at least may register the difference between their own and the other's perspective, in some manner, in such cases, even if the task they are doing occupies much of their attentional resources. Yet, this effect of the other's underspecified belief is either weaker than the effect of specified beliefs or cannot be reliably captured via measuring where attention is directed.

3.7. General Discussion

The aim of the present study was to investigate whether people spontaneously represent the alternatives other agents, who are uncertain about the actual state of affairs, presumably do, even though such contents are complex (i.e. are made up of multiple elements, that are connected by a logical operator) and may not render others' behaviour immediately predictable in a situation. To this end, we designed a task in which a human avatar, who either did or did not witness the hiding of a ball, could entertain two 'hypotheses' regarding the ball's location when he did not saw the hiding, and measured whether in this situation participants demonstrate an attentional bias towards both locations the agent considers a 'possible' hiding place. In specific, we investigated whether participants' sensitivity to detect changes at the location that was empty but was considered a potential hiding place by the other, is facilitated in situations where the other agent represents two equally likely alternatives (specifically the disjunction: '*the ball is either at location A or at location B*'). We tested this research question in experiments where the other agent's belief was irrelevant for the task and when participants had to monitor the agent's belief regarding the location of the ball.

Throughout five experiments we found strong and consistent evidence that participants' own representation of the actual state of the world (the ball's location) facilitates their performance in our change detection paradigm, i.e. evidence that our measure is capable of capturing the impact a represented content on the spatial attention of human adults. These results corroborate and extend previous findings which demonstrated a similar effect of adults' own representation of the state of affairs on their attention, specifically on their anticipatory looking behaviour, in studies using the visual world paradigm (see e.g Altmann & Kamide, 2007), by showing that such attentional effects emerge even in the absence of a linguistic input or the expectation of the appearance of an object. Furthermore, the results of our last experiment indicate that the hypotheses one entertains (regarding the location of an object) can have a similar impact on spatial attention as factual knowledge, suggesting the format and the external validity of the represented content may not matter in whether or not its effect emerges in where one's attention is directed at a certain moment.

Importantly, in Experiment 2, we also demonstrated that our paradigm can capture not only the effect of one's own, 'specified' and 'underspecified', but also the effect of another agent's 'underspecified' belief content, on how people allocate their spatial attention among objects in their environment. In particular, we demonstrated that if participants deliberately monitor the content of another agent's hypothesis space (what the agent may think 'possible' in the given situation), because they are

instructed to do so, this voluntarily computed content affects their sensitivity for changes at the locations corresponding to the hypotheses entertained by the other. This finding provides further evidence that once the content of another agent's mental state is computed, it affects people's own representation of the state of affairs in an uncontrollable manner, even in tasks and situations where the content is irrelevant. This, in turn, lends further support to the view that the representation of others' mental states is not completely decoupled from one's own representation of reality (or self-knowledge), either because the content of the self- and the other-perspective are stored in parallel or the two are linked to each other in some manner (Perner et al., 2015; Perner & Leahy, 2016). Crucially, it also suggests that the effect of other agent's mental state on human adults' behaviour, may be independent of both the form and the complexity of the attributed content. Notably, the effects observed on the underspecified belief trials – both in this and in our last experiment – were present from the beginning of the task, which excludes the possibility that the attentional bias, the reaction time profile indicated, was merely the result of an associative learning process (between the hiding, as an event, and the size/surface features of the boxes), that eventually led to the 'highlighting' of the two large boxes.

Despite all these findings, proving the validity of our paradigm, in three of the four experiments where the other agent's belief was irrelevant and the ball was present at the timepoint of the measurement, we did not find the predicted effect. In specific, we did not find a significant group-level difference between our two critical conditions along our reaction time measure, thereby not finding firm and convincing evidence for the spontaneous representation of the other's hypotheses. Crucially, when the ball was removed from the scene, that is, we eliminated the well-known 'pull-of-the-real', the predicted difference emerged, i.e. participants were faster to detect changes at the location the other agent could consider a potential hiding place (compared to the impossible locations), though the effect was small, and was present in only about half of the sample. Even though the replicability of the finding is somewhat questionable, and it is not entirely clear what the great individual-level differences reflect (a state- or trait-like feature of spontaneous mentalizing, e.g. a general tendency to take the perspective of others), this result suggests that human adults *may* represent the alternatives other agents consider, spontaneously, provided that the attentional demands of the situation do not exceed a certain level (as it was the case in, for instance, all those experiments of Study 2 where participants had to override the pull-of-the-real to be able to attend to all four boxes). Such situations may arise, for instance, when the observer does not have a firm knowledge about the state of affairs, or when his/her knowledge is not relevant in the given context (as it was the case in the 'ball absent' condition, where there was no need to track the actual location of the ball). The results of Experiment 4, namely the fact that evidence for representing the other agent's underspecified belief content was present on the 'ball absent' but not (or less clearly) on the 'ball present' trials, corroborate the crucial role of the

availability of executive resources in the spontaneous computation of the other agent's visual perspective/belief content, which has been demonstrated in a number of previous studies that investigated the impact of cognitive load on the tracking of other agents' mental states (see e.g.: Schneider, Lam, et al., 2012; Qureshi & Monk, 2018). They also show that the role of the individual's executive resources (in the tracking of others' beliefs) may scale with the complexity of the to-be-represented mental state content.

Importantly, the lack of the predicted effect in those experiments where the ball remained present, after hiding in one of the boxes, does not mean that in these studies participants did not track the knowledge state of the other. The pattern of miss ratios suggest that they differentiated the condition when the agent knew where the ball was and when he was uncertain about where it has hidden, at least in some experiments, with some results indicating a less strong 'reality bias', others an enhanced attention towards the 'possible' location of the ball, on those trials where the agent did not witness the hiding event. These results imply that participants may have actually encoded the content of the avatar's belief, in some manner, it just did not influence the speed of their reactions to changes at the 'possible' location, either because the actual location drew too much attention and no resource was left to monitor the 'possible' box, or because, counteracting the reality bias generated relatively high cognitive load, under which participants may have merely represented the other agent's ignorance (but not the actual alternatives he considered).

Taken together, our findings suggest that, while human adults may represent other agents' hypotheses, spontaneously, at least under certain circumstances, our measure might not be sensitive enough to reliably capture the potentially very small modulatory effect of the involuntarily computed content. This is not very surprising if we consider how many factors may affect where attention is directed at a given moment, from low-level ones, like the spatial position or the salience of an object, to high-level ones such as, for example, whether a certain location was associated recently with a task-relevant event (such as the hiding on the object). To exert its influence on participants' spatial attention, the content of another agent's underspecified belief has to override the effect of all of these factors. It might succeed in some cases but not in others, resulting in a large individual heterogeneity, both in the presence and the magnitude of the effect, as well as a huge fluctuation, both across trials and experiments. On this account, the large individual differences observed in our experiments reflect the varying sensitivity of participants' attention to the attributed content, which may be a trait-like feature of human adults' cognition or depend on the availability of executive resources at the given moment, the other agent's mental state content has to be computed. Further studies need to elucidate whether (or rather: to what extent) the observed large inter-individual variability reflects simply differences in the sensitivity of attention to the represented content or (i) differences in the

participants' general 'other-directedness', (ii) propensity to spontaneously represent the mental states of others, (iii) capacity to inhibit their own perspective.

Given the limitations of the paradigm, future experiments, addressing the issue whether human adults represent the content of other agents' hypothesis space in a spontaneous manner, should use different, possibly more sensitive measures, that are less susceptible to the effect of low-level factors than spatial attention, and/or tasks in which the other agent's belief has more relevance for the participants. Alternative approaches include examining the capacity in interactive settings, investigating human adults' expectations regarding how another agent will search, after, for example, he/she could exclude certain possibilities or testing the ability to track others' hypotheses about properties of objects that go beyond the location they occupy at a certain moment, such as their function, their affordances, or causal role in certain events.

Chapter 4: Tracking other agents' inferences

4.1 Theoretical background

Throughout the four decades of ToM research, most of the studies exploring people's understanding of others' minds have investigated situations in which the content of the other agent's belief was determined by what events he/she has witnessed (Kovács et al., 2010; Wimmer & Perner, 1983; Schneider et al., 2017) or has been informed about by others (Király et al., 2018; Song et al., 2008; Tauzin & Gergely, 2019) and tested whether participants can correctly anticipate what the agent will do next. Importantly, however, people's actions are guided not only by beliefs formulated on the basis of the input directly available at a given moment but also based on beliefs generated via reasoning. To be able to prepare for the potential actions another person might execute in the future, people must not only encode what other agents are aware of at a certain moment but also have to consider what inferences others may draw from the beliefs they may hold. This enables people to generate long-term predictions regarding what they should expect from others and thereby flexibly adapt to possible changes in their behaviour. For instance, when a burglar tries to predict the next move of the detectives, to be able to mislead them and thereby avoid being arrested, he has to consider not only whether the detectives have noticed the clues left at a crime scene (open door with no sign of breaking in), but should also represent what these clues mean to them, i.e. whom they may suspect, based on the evidence they have, given their situation-specific knowledge and stable beliefs that potentially affect their reasoning process (*'the burglar is someone who has a key, therefore it is either the ex-boyfriend or a close relative'*).

Human adults are clearly able to perform such computations deliberately, when they consciously invest effort to do so, both offline and while being involved in an interaction. They can follow the logic of their opponents in strategic games, where the goal is to outwit each other, or the line of thought of people they are trying to deceive, like the burglar in the example above, as well as the argumentation of their partners in discussions. They easily understand such interactions, as the audience of crime series, open debates, and TV shows, understanding why different characters, whose decisions are often motivated by the inferences they have drawn, act the way they do, without even realizing that they are following their line of thought. This suggests that human adults might track others' inferences spontaneously, without even having the intention or being aware of doing so. Whether this is indeed the case and human beings are endowed with such a capacity, as part of the social-cognitive repertoire that enables them to function efficiently in the social world, is an issue yet to be explored. The main aim of the present study is to investigate this issue.

Although important in its own right, the question has special significance also for another reason: it has outstanding relevance for the ongoing debate about the representational underpinnings of the ToM abilities that underlie online social interactions in adults and the performance of infants on

nonverbal ToM tasks. Genuine understanding of others' minds entails the full understanding of the functional role of beliefs: that perceiving an object or event normally leads to a true belief about it, that beliefs lead to decisions that result in actions, and, crucially for us, that beliefs are inferentially integrated, that is, they lead to other beliefs via reasoning, as well as to decisions in combination with other mental states (Rakoczy, 2012). This means that a fully-fledged theory of mind should enable the tracking of both types of inferences humans routinely make: 1. inferences about what belief one may form based on the available input, and 2. inferences from one belief (p) to another (q), using some kind of logical rule (e.g. if p then q; p; therefore q). Besides these, as traditionally argued, it should also enable inferences from the other's belief (or the belief and the other's goal) to the output, i.e. to behaviour. Accordingly, if human adults can and do represent the beliefs of others spontaneously, they should be able to track the inferences of others, in a similar manner, regardless of their type. That is, they should be able to spontaneously represent not only what inferences another person may draw from perceptual evidence (e.g. if they see their train leaving from the target platform, they will believe that they have missed it) but also other types of inferences, including the logical inferences agents presumably perform in a situation (e.g. if they see two trains on the target platform, and learn that the one leaving is not their train, they will infer that the other train must be theirs). Uncovering whether people track such inferences spontaneously, may have important implications for the long-standing debate, discussed in Chapter 1 (see e.g. Butterfill & Apperly, 2013), whether representational flexibility and efficiency can coexist in mental state attribution, i.e. whether it is possible to attribute complex mental state contents quickly and efficiently during online social interactions.

Notably, representing other agents' logical inferences does not necessarily mean representing the reasoning process itself the other goes through. People may rely on their own reasoning mechanism when they engage in such computations (whether they do it deliberately or not), just feed in the information available for the other agent instead of what they know (in case the others' knowledge differs from their own). Representing others' logical inferences thus essentially entails: (1) representing the beliefs the other presumably holds and (2) carrying out the appropriate inferences on those. The processes involved in such inferences may largely correspond to those one recruits when performing inferences from first-person perspective. Crucially, however, at the end of the inferential process performed from a third-person perspective, the other's mental state content gets updated by the resulting conclusion and not observer's own.

Importantly, evidence suggests, that some of the logical inferences one makes from a first-person perspective, in a non-social context, may take place spontaneously in humans, opening the question whether this is also the case when such inferences are made from a third-person perspective. Human adults perform certain, elementary logical inferences, such as the modus ponens *(„if p then q"; 'p"; „therefore q"*) or the disjunctive syllogism *(„p or q"; „not p"; „therefore q"*), with high success rates,

without any formal training in logic, both in daily life and in experimental settings (Evans et al., 1993, as cited in Reverberi et al., 2012; see Reverberi et al., 2007 for similar results). According to one view, they do this by relying on a set of inferential rules or schemas, which define what kind of conclusion can be drawn from premises of a given form (Braine et al., 1995). According to an alternative view, they construct mental models to represent the premises which are then combined in turn (Johnson-Laird & Ragni, 2019; Ragni & Johnson-Laird, 2020). Both accounts agree, however, that deductive reasoning is a multi-stage process that requires the integration of the premises that are first represented (and maintained in working memory) separate from each other, and that some elementary deductive inferences should be 'easy' compared to others, implying that they should take place spontaneously in humans.

While for modus ponens evidence clearly suggests that this is the case, in line with what both theories suggest (as this type of inference requires the application of only one inferential rule or, using the terminology of the mental model theory, the construction of only one initial model), for disjunctive syllogism the results are more mixed. For instance, in a priming task, in which adult participants were first presented one premise supraliminally (e.g. 'if 2 then 5') then another one subliminally (e.g. '2') and finally had to judge whether a number was even or odd, Reverberi and colleagues (2012) found that, in case of conditional statements, participants were faster to perform the judgement when the presented number matched the conclusion following from the two premises (e.g. was '5'), compared to when it did not, suggesting that they performed these inferences automatically, upon receiving the second premise. In contrast, no such priming effect emerged when the first premise was a disjunctive statement, indicating that disjunctive inferences may be computationally more demanding than modus ponens.

Developmental studies which find that the ability to represent disjunctions and reason by exclusion develops only by the preschool years, provide further support for this proposal (Gautam et al., 2021; Mody & Carey, 2016). In specific, using a complex version of the so-called 'cups task', Mody & Carey (2016) found that if children under age three witness the invisible hiding of two rewards in two pairs of cups (one in each pair), and then learn that one cup is empty, they do not tend to search more in the other cup belonging to the same pair, indicating that they do not yet understand the dependent relationship between the alternatives (that if one alternative turns out to be false the other one *must be* true and vice versa), a crucial component of disjunctive reasoning. In contrast, more recent studies, using simpler scenarios, with only one 'unknown variable' (the identity of one object) and two hiding locations (instead of four) and a different way to measure whether participants perform disjunctive syllogism, suggest that adults and even 12-month-old infants perform such inferences spontaneously, when this is necessary to interpret events that are ambiguous. In specific, their pupil dilation, as well as the looking time patterns of infants suggest that adult and infant participants spontaneously apply

disjunctive syllogism when it is possible to resolve ambiguity regarding the identity of a hidden object by doing so (Cesana-Arlotti et al., 2018; Cesana-Arlotti et al., 2020). This indicates that, unlike what the above-mentioned adult and developmental studies suggest, performing such computations, in itself, may not demand substantial effort. Furthermore, other results imply that, for infants, it actually may not be more difficult to encode the preference or the goal of an agent based on what they could infer by disjunctive syllogism than to represent it based on perceptually available information (Cesana-Arlotti et al., 2020). This raises the possibility that the ease with which the content of another agent's mental state is computed may be independent of whether the person performing the attribution makes inferences based on what the other person has direct perceptual access to or adheres to logical inferences.

Notably, the spontaneous tracking of the logical inferences other agents may perform in a certain situation requires not only the ability to readily perform deductive inferences from first-person perspective and the ability to continuously track and take into account what the agent is aware of. As mentioned earlier, the observer also has to (1) maintain the content of the represented mental states (that can serve as „premises" for the subsequent inferences) in the working memory or in some domain-specific buffer (e.g. [he believes that] '*the burglar is either a stranger, a close relative or the ex'; 'as the door was opened with a key the burglar is not a stranger'*) until premise integration can occur and (2) update the previously attributed belief with the conclusion ('*therefore it is either the ex or a close relative'*). In case of more than two alternatives, such an update may take place more than one time until the person arrives at the final conclusion[19].

It is important to point out, that, despite the apparent complexity of the above-described process, some of these computations, in particular the updating itself, may not be cognitively demanding. Recent developmental studies suggest that, even 13-month-old infants revise the content of attributed beliefs spontaneously, when the other agent receives a new piece of information (Song et al., 2008; Tauzin & Gergely, 2019), which indicates that humans are able to perform manipulations on the attributed mental state content from relatively early on, and do so readily, when necessary, to be able to interpret and predict others' actions.

Taken together, evidence suggests that humans may possess all the abilities necessary for tracking the logical inferences of others: they readily perform deductive inferences, from a first-person view, and

[1919] Given the multi-step nature of the process one might argue that it is unlikely that such computations would take place spontaneously. Indeed, if the inferential chain involves several steps and/or the person has to take into account the other agents' different knowledge state throughout the process, the representation of others' logical inferences may place heavy demands on executive functions. Without investing much effort, there is a rather high chance of feeding in the wrong premise at one point (for instance, conclusion the person himself could draw from the available information, instead of the agent's conclusion) and thereby confusing what one can conclude with what the other might have inferred. Here we investigate simplier cases, in which the number of inferential steps one has to go through is relatively low, aiming to avoid cognitively too demanding situations.

spontaneously update the content of other agents' mental states from an early age. Hence, theoretically, there are reasons to assume that they are also able to represent what conclusions another agent can draw from the beliefs she holds in the same manner, i.e. involuntarily and without much effort. They should be able to do this at least in those cases where the inference requires the use of only one or few rules (or the construction of a few mental models), and, as such, would be considered 'easy' also when performing the computation from first-person view. Results of a developmental study, showing that 14.5-month-old infants can attribute erroneous conclusions to other agents based on a misleading piece of information - where a doll is, based on a protruding material that resembles part of the doll (Song & Baillargeon, 2008) -, imply that the ability to represent other agents' conditional inferences, based on what they have perceptual access to ('*if blue tuft [is protruding] then it is the doll*'), may be present from very early on. Nevertheless, since the study was not designed to test this ability, and one may argue that conditional inferences may not even be necessary in this situation (the observer could, for instance, infer that the other will mistakenly believe that he sees the doll at the particular location, taking the protruding part as direct evidence for the doll's presence at the given location), it is unclear whether the finding can be considered as an evidence for tracking others' deductive inferences in infants. Given the scarcity of experimental studies targeting the issue, it remains an open question whether humans represent the conclusions other agents may draw from what they believe or know spontaneously (i.e. without being guided by a specific intention or external prompt) and more generally, how they track the inferences other agents draw.

The present study aimed to address this issue. It had two main goals: first, to investigate human adults' ability to represent the logical inferences of other agents; and second, to test a specific assumption regarding the nature of this ability, i.e. that adults track such inferences spontaneously. In specific, we aimed to investigate (1) whether adults are able to track what conclusion another agent may draw from the beliefs she holds (on the basis of what events she witnessed or what she has been told) and (2) whether the other's (potential) conclusion modulates adults' task performance, namely their own estimations about the likelihood of certain events and the time necessary to perform such estimations, even if they do not have the intention to track the belief of the other. We tested whether a mismatch between the participant's own conclusion and the conclusion another agent can draw in the same situation (but based on different beliefs), generates a similar 'altercentric interference effect' that has been observed in a number of implicit ToM tasks where the other agent had an incongruent visual perspective or held a divergent belief (Kampis & Southgate, 2020). In particular, we focused on a situation, where the observed agent ends up representing two alternatives, for instance, regarding where an animal has hidden, while the observer represents only one. Consequently, the agent considers an option the participants could exclude from the range of options, as 'possible'. We tested whether participants rate the probability of this alternative higher than the probability of the

alternative both of them consider 'impossible' not only when taking the other agent's perspective- but also when performing the task from their own perspective.

To this end, we designed four experiments. Experiment 1 tested whether human adults represent the alternatives another agent presumably does regarding the location of a hidden object, spontaneously, when such a belief content can be inferred from the events the agent witnessed by applying disjunctive syllogism. Participants were presented with sequences of pictures displaying a girl and three boxes. They were told that a kitten has hidden in one of the boxes and their task was to judge how likely it is that the animal is hiding at a specific location, after watching the scenario. During the trials they only received indirect evidence regarding the kitten's location, specifically, they were shown that two of the boxes were empty. Thus, the kitten's location could be non-ambiguously identified by applying disjunctive syllogism *('it is neither in Box1 nor in Box2, therefore it is in Box3'*). Both the agent and the participant witnessed the opening of the first box. Our crucial manipulation involved whether the agent witnessed the second box opening (true belief – 'TB trials') or not (underspecified belief – 'UB trials'). All this time the third box remained closed. At the end, participants had to rate the likelihood that the kitten has hidden at a certain location either from self-perspective (on trials that started with a YOU cue) or from the perspective of the other agent (on trials starting with the SHE cue), on a continuous scale. The to-be-rated location was either the (i) kitten's actual location, (ii) the box that opened first and was known to be empty both by the participant and the agent ('impossible' location) or (iii) the one that opened second, hence, on the underspecified trials, could be excluded from the range of options by the participant but not by the agent ('possible' location). We hypothesized that if participants spontaneously represent the other's inferences, we should observe an altercentric bias on those trials where participants make judgements about this 'possible' location from first-person perspective, i.e. a tendency to estimate the likelihood of the kitten being at this location higher than the likelihood of the kitten being at the impossible location.

In a certain sense, Experiment 1 investigated the same theoretical question we did in Chapter 3, specifically, whether adults represent other agents' underspecified beliefs spontaneously, for which failed to find convincing evidence in Chapter 3. Importantly, however, in the experiments reported in the present chapter, we used a different paradigm and a different measure, about which we assumed that they have more potential to detect altercentric interference effects, in general, and the spontaneous representation of other agents' hypotheses, in particular, for a number of reasons, listed as follows. First, the current design and the way information was presented to the participants resembled more that of the classic „reasoning by exclusion" tasks, which recently provided convincing evidence for the spontaneous representation of multiple, mutually exclusive alternatives, from first-person perspective (Cesana-Arlotti et al., 2018; Cesana-Arlotti et al., 2022). Second, by demanding continuous switching between the self- and the other-perspective and providing perspective cues on

every trial, the task also resembled more the spontaneous visual perspective-taking tasks, which were the first to demonstrate 'altercentric intrusion' of the other's perspective into one's own (see e.g. Samson et al, 2010, Experiment 1) and the results of which were replicated several times later (see Kampis & Southgate, 2020). Both of these aspects of the design may highlight that the agent holds a different belief in some cases (due to not being able to eliminate one of the alternatives). Third, since in the current task the content of the agent's belief (e.g., that the kitten is either in Box A or Box B) matched what participants had to represent on the self-perspective trials to be able to perform well (the potential location of the kitten), it could be expected to exert more influence on participants' responses than the content of the agent's belief did in the change detection task we used in the experiments presented in the previous chapter, where there was no such match. Finally, we assumed that likelihood ratings, provided on a continuous scale, may better capture the potentially small modulatory effect of the other agent's belief content on participants' own representation of the state of affairs than the measure we used in the previous chapter as they provide opportunity to detect subtle differences, and, at the same time, may not be influenced by as many low-level factors as decisions based on spatial attention (such as the spatial layout or the surface features of objects, that exert strong impact on where attention is directed).

Experiment 2 was a conceptual replication of Experiment 1, in which participants (and the agent) had to infer the identity (instead of the location) of the hidden object. Experiment 3 tested a more complex case in which the final conclusion could not be drawn without first performing another inference and where participants had to combine of two logical rules to identify which animal has hidden in the scene. Finally, Experiment 4 investigated a situation in which participants also had to make multi-step inferences and combine two logical rules but the attribution of the appropriate content required taking into account a stable, situation-specific belief of the other agent (e.g. she believed an animal can hide both in Box 2 and Box 3 while the participants knew that it can only hide in Box 3), instead of continuous tracking what she had and had not witnessed, on a trial-by-trial basis.

Crucially, as mentioned above, in all experiments, trials started with a prompt that defined whose perspective participants had to take (self or other). Such a manipulation rendered tracking the girl's belief on self-perspective (or 'SELF') trials unnecessary. We hypothesized that, if participants track others' logical inferences *explicitly*, when they have to judge the likelihood that the animal has hidden at the presented location or that the presented animal has hidden in the scene from the agent's perspective ('OTHER' trials), then, on those trials where the other agent cannot be certain about the animal's location/identity, they should provide similar ratings for location where the animal actually hides/the actually hidden animal and the one that is 'possible' for the agent. Importantly, regarding our main hypothesis, if they also do this spontaneously, even when they do not have to track the agent's beliefs, this should bias their estimations on the possible trials in the SELF 'underspecified

belief' condition, when the agent represents two alternatives (but not on the true belief trials where she represents only one, the same as the participant). In specific, participants should provide higher ratings for the possible compared to the impossible alternative (or, in Experiment 4 for the animal linked with the box about which the agent has a mistaken belief) and/or it should take them longer to perform these ratings on these 'possible' trials, indicating interference from the other's perspective. Depending on the limitations of this ToM ability – whether humans can spontaneously track more complex, multi-step logical inferences or can represent the conclusions others may draw spontaneously only in simple cases - this pattern should be present either in all or only in those experiments, where the inference participants had to perform is made up of only a few steps and the final conclusion requires the use of only one inferential rule (Experiment 1 and 2).

4.2. Experiment 1

Experiment 1 tested whether human adults represent the alternatives another agent may uphold regarding a hidden animal's location, when the agent's belief content has to be inferred from what events she witnessed via disjunctive syllogism. Importantly, on certain trials the agent had access to less information than the participant, in specific, contrary to the participant, she did not see the content of the second box. Thus, the conclusion the agent could arrive to (e.g. *'the animal is either in Box1 or Box2 or Box3', 'it is not in Box1' 'therefore is either in Box2 or Box3'*) differed from the one the participant could draw *(e.g. 'it is neither in Box1 nor in Box2, therefore it is in Box3'*). Specifically, we tested whether participants perform these inferences spontaneously or only when they are instructed to do so.

4.2.1. Methods

4.2.1.1 Participants

Participants were 35 adults (M_{age}=29.17, SD_{age}=4.06; 20 males), recruited via Testable Minds, an online platform to recruit participants for behavioural and psychological experiments (minds.testable.org). Selection criteria included: English as a first language, age below 35 years and >70% approval rate on previous studies as well as UK/USA/New-Zealand or Australia listed as their current location, to ensure that they understand the instructions that were in English and good enough data quality. All of them were members of the Testable Minds pool, i.e. they had their location and identity verified and were authenticated with Face ID. All but three of them were right-handed and all of them had at least a high school degree. The target sample size was determined a piori, using G*Power 3.1. Assuming a medium effect size (d=0.50) a sample size of 36 participants provided an 80% statistical power for a two-sided paired-samples t-test, with an alpha of 0.05. One additional participant was tested but her data was excluded from the analyses because the ratings on the SELF actual and/or the SELF impossible trials indicated lack of understanding of the task or the use of the rating scale. Further 20 participants were tested but were not included in the analyses: 7 because they had less than 70% valid trials, 11 because they failed to answer >50% of the attention check trials correctly, and 2 because they failed to meet both criteria (the exclusion and inclusion criteria see Section 4.2.1.5: Data analyses). The study was approved by the EPKEB United Ethical Committee, Hungary; participants gave informed consent prior the experiment and received monetary compensation of 5.3 USD for their participation, via the Testable Minds platform.

4.2.1.2 Stimuli

Stimuli consisted of a set of images (1152 x 648 pixel size each), and a grey rating scale on which participants' responses were recorded. The images depicted a simple environment with a girl in the background and three coloured boxes of the same size at the front, approximately equal distance from the girl: a blue on the left, a yellow on the right and a red at the centre (see **Figure 4.1a** for an example). The position of the three boxes was kept constant throughout the task. The boxes were either all closed

on the pictures or one or two of them were open, such that it was obvious that they are empty. The latter two types of pictures had two versions: one, on which the girl was facing the boxes, thus she had the same knowledge as the participant, and one on which she turned her back towards them, hence she had no visual access to the piece of information revealed on the respective picture. The scene and the boxes were created in Microsoft Powerpoint, the clipart image of the girl was downloaded from the internet (https://www.dreamstime.com/) and was modified in Adobe Photoshop CS6.

The grey rating scale was a 750 x 146 pixel image made up of two parts: the grey version of a colour gradient scale (originally ranging from red to blue), extending 750 pixel in length and 85 pixel in width, and two arrows above the scale, pointing to the left and the right from the middle, with the expressions *less likely* and *more likely* written on them (see **Figure 4.1c**). To prevent participants from using only the endpoints of the scale, those were not labelled in any way. All stimuli were displayed on a plain white background, in their original size.

4.2.1.3 Apparatus

The experiment was built and hosted on Testable (www.testable.org), a website for creating and hosting online behavioural and survey-based studies. Participants used their own computers and internet browsers, and either a mouse or a touchpad, with the browsers restricted to Chrome or Firefox and devices to PCs and laptops on the hosting website to reduce variability resulting from the different softwares and hardwares used by the participants. They were also asked to use a minimum 1280 x 720 pixel screen resolution to ensure that all pictures can be displayed in their original size. Active screen area sizes ranged from 1280x720 pixels to 2144 x 1206 pixels (data automatically collected by Testable).

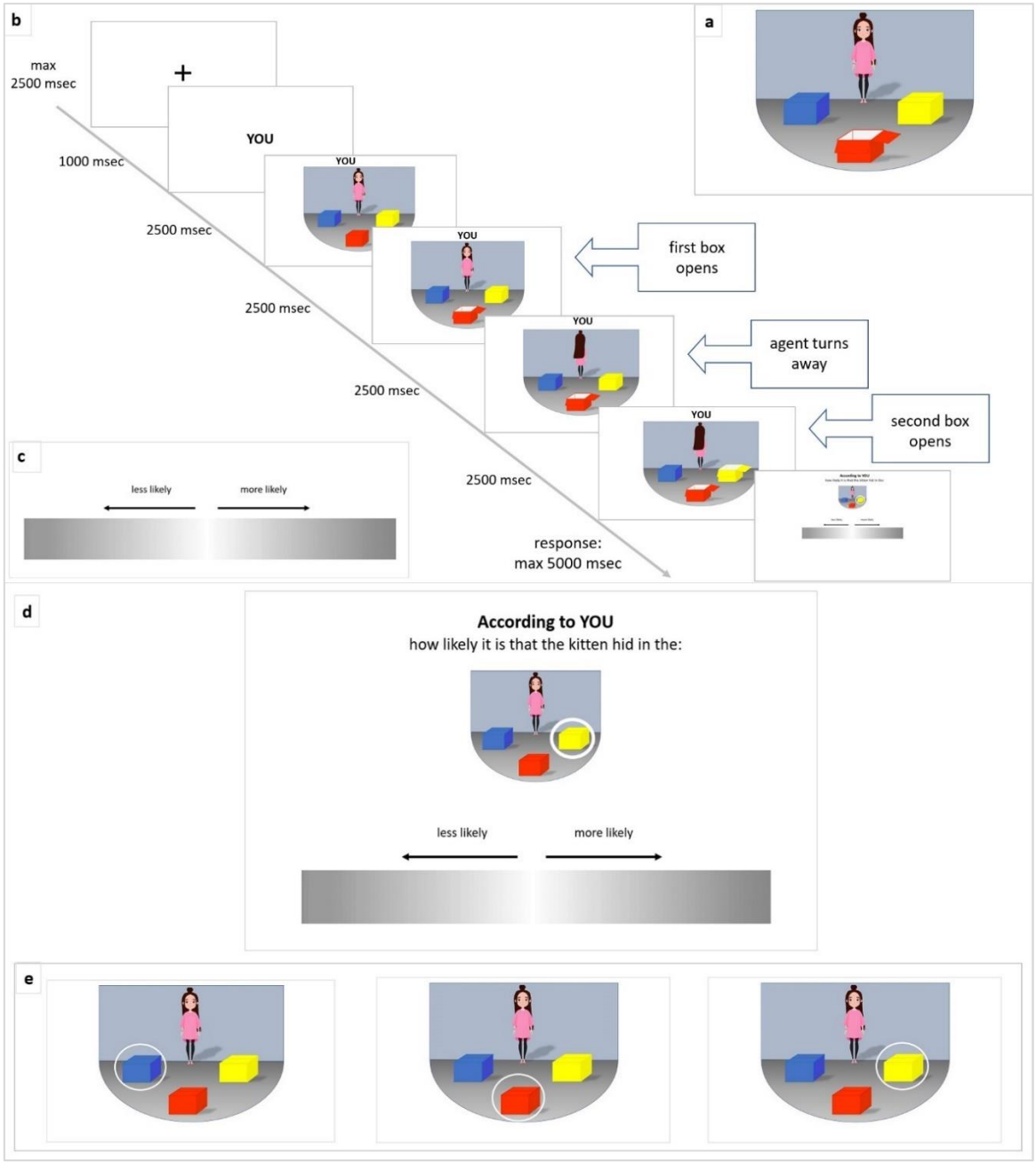


Figure 4.1. (a) An example of the pictures presented in Experiment 1. (b) The trial structure of the main task. Participants saw picture-sequences on which first one then another box opened with an agent either witnessing the second box-opening (true belief trials) or not (underspecified belief trials). They had to estimate the likelihood that a kitten is hidden at a certain location indicated on the picture presented either from their own (YOU prompt; SELF trials) or from the girl's perspective (SHE prompt; OTHER trials). The location could be the kitten's actual location, an impossible one (the box that opened first and was known to be empty both by the agent and the participant) or a possible location (which could be considered a potential hiding place by the agent on the underspecified belief trials). The example presents a SELF underspecified belief, 'possible' trial. (c) The rating scale on which participants had to indicate their responses. (d) The response screen presented at the end of the main task, with (e) the three possible to-be-rated locations below (from the left to right: the actual, the impossible and the possible alternative, in the light of the even sequence depicted on Figure 4.1b).

4.2.1.4 Procedure

After providing consent to participation participants were asked first to perform a built-in calibration procedure of Testable, to ensure that all images would appear in their original size, irrespective of the participant's screen resolution. Then they were instructed to set their browser window to full screen. Following this, participants provided demographic data (on age, sex, education and nationality). Finally, to guarantee that we would have the x-coordinates of the endpoints of the scale, necessary for the analyses, participants performed another calibration: they were presented with the image of the rating scale and were asked to click first at the leftmost then at the rightmost point of it, marked with a red fixation cross. After this, the experimental session started, with a general instruction screen explaining the structure of the task.

The experiment was made up of three main phases: two short training sessions made of 6-6 trials, and a main task, comprising four practice and 72 test trials. Each phase was preceded by its own instruction screen (see Supplementary Materials **S4.1.1**), with a maximum time limit set, to prevent participants from disengaging from the task for longer time periods[20]. The experimental session ended with questions asking participants: 1) how difficult they found the task 2) whether they had any idea about the purpose of the study 3) and whether they used any specific strategy during the task. The whole task lasted for about 30-35 minutes.

4.2.1.4.1 Training sessions

Training session 1

The purpose of the first training session was to demonstrate participants how an open box with or without a kitten inside looks like and thereby guarantee the appropriate understanding of main task (that only the closed box can contain the kitten). To this end, we presented participants with pictures depicting two open and one closed box, with either a partially visible kitten in one of the open boxes (‚kitten present' condition) or two empty boxes ('kitten absent' condition). On each trial participants

[20] The time limits were piloted to make sure that it is long enough to allow reading of the presented text and processing of all the information.

first saw a central black fixation cross (85 pixel in height and 95 pixel in width), for 500 ms, then a blank screen for 250 ms. Pictures were presented after this interstimulus-interval (ISI) for a maximum of 2000 ms or until a response was provided. Participants' task was to decide whether either of the open boxes contains the kitten and click on the box that does, as fast as possible (or do nothing if both boxes were empty). Following their response and a second 250 ms ISI they were provided with written feedback, which remained on screen for a fixed 2000 ms. Then, after a 250 ms intertrial-interval the next trial started. Participants received altogether six trials: two 'kitten absent' and four 'kitten present' trials, in a pseudorandom order. Since the training session mainly served a demonstrative purpose, progress was independent of performance.

Training session 2

The second training session was included to ensure that participants interpret the scenes correctly, that when the agent sees one box open (and empty) and does not witness the second box opening she cannot know in which of the remaining two boxes the kitten has hidden. To this end, participants were presented with picture-sequences similar to those of the main task, displaying a girl and three boxes, of which first one then another one opened, with the girl either witnessing or not the second opening. Participants were told that the kitten has hidden in one of the boxes and, in the end, they would have to decide whether the girl would search for it in one or two boxes.

Each trial started with central black fixation cross (85 pixel in height and 95 pixel width), which was presented for 1000 ms. Following this, four pictures were presented sequentially, each for 2500 ms, at the centre of the screen, with no interstimulus interval, to ensure the smooth unfolding of events. The first picture was always the same, depicting three closed boxes and the agent that was facing them. The second and the third picture displayed one, and the fourth (and last) picture two open boxes. Box openings (on the third and the fourth picture) were accompanied by a sound to make sure that participants encode these events. Importantly, on the third and the fourth picture the girl was either facing or was with her back towards the boxes, yielding two conditions: one in which the girl could infer the kitten's actual location ('true belief' condition) and one in which she could not be certain about it, i.e. represented two and not just one alternative (underspecified belief' condition). Finally, after the presentation of the fourth picture and a subsequent 250 ms ISI, a blank screen appeared with the question *'Would the girl search for the kitten in?*' at the centre and two buttons below it, on the left and the right, with the 'ONE BOX' and 'TWO BOXes' expressions displayed on them. The appearance of this screen marked the beginning of the response period, which lasted for a maximum of 5000 ms or until a response was provided. Participants had to click on the selected button as fast as

possible. Responses were followed by feedback. In case of an incorrect response participants received a text indicating the correct button and providing an explanation why that button would have been the correct choice. Progress was self-paced to leave enough time for processing the presented text. Altogether participants received four underspecified belief and two true belief trials, in a fixed pseudorandom order, with each possible open box combinations (blue-red, blue-yellow, yellow-red) presented twice during the session. Crucially, in case they failed to respond correctly on more than two underspecified belief trials or on the two true belief trials, the session was repeated once.

4.2.1.4.2 Main task

The trial structure of the main task was similar to that of the second training session with the key difference that participants rated the likelihood that the kitten has hidden in a certain box (the one encircled on the presented picture) at the end, either from their own or from the girl's perspective and received no feedback for their responses (see **Figure 4.1b**). Each trial started with the presentation of a central black fixation cross, 85 pixel in height and 95 pixel in width. To fix the position of the cursor on the screen and thereby ensure that it will reappear later in the middle of the rating scale, participants were instructed to click on the cross (exactly where the two lines cross each other), on each trial, as fast as possible. They had a maximum of 2500 ms for this, after which the fixation cross disappeared. Following this, participants were presented either with the word SHE or YOU at the centre of the screen which indicated, in advance, whose perspective they will have to take when they will be asked to estimate the likelihood that the target animal is at a certain location at the end: their own (SELF perspective trials) or that of the girl (OTHER perspective trials). Following the perspective prompt, they saw a four-picture sequence, made up of the same elements and presented in the same way as the sequences presented during the second training session, with the girl either holding a 'true' or an 'underspecified' belief regarding the kitten's location by the end. Importantly, to guarantee that participants take the appropriate perspective while watching the events, the perspective prompt remained visible (above the pictures) throughout the whole sequence. Finally, after the last picture and a 250 ms ISI, a blank screen appeared with the *'how likely it is that the kitten is in the*' question at the top, a small version of the first picture, with one of the three closed boxes circled on it below and the image of the scale at the centre. Depending on the perspective participants had to take in the trial, the question was either preceded by the *'According to the GIRL'* (OTHER trials) or the *'According to YOU'* (SELF trials) expression. The appearance of the scale marked the beginning of the response period

for the participants, which lasted for 5000 ms or until response was provided. Trials were followed by a 250 ms intertrial-interval, during which the screen remained blank.

Depending on the location participants had to rate, trials could belong to one of the following three experimental conditions: actual (if the target of the question was the actual location of the kitten that could be inferred), impossible (if the question was about the box that opened first) or possible (if it was about the box that opened second, thus, on the underspecified belief trials, could be considered as a potential hiding location by the girl). Thus, the experiment followed a 2 (SELF or OTHER perspective) x 2 (belief: true or underspecified) x 3 (alternative type[21]: actual, possible, impossible) design.

Participants first performed four practice trials (two SELF and two OTHER perspective trials, with one true and one underspecified belief trial for each perspective) without feedback, to familiarize them with the structure of the test trials. Following this, they received 72 test trials, in two blocks: 18 per perspective and belief type, with each type of alternative presented six times within each perspective and belief combination. The two blocks were separated by a short break. Participants were encouraged not to leave the screen and continue within 3 minutes, but progress was self-paced.

The perspective, the belief, the alternative type and the identity of the to-be-rated box were counterbalanced within blocks as well as which box opened first. The order of the trials was pseudorandomized within the blocks, such that there were no more than three consecutive trials with the same perspective, belief, alternative type, to-be-rated box, and sequence starting with the opening of the same box in a row. Six different trial orders (i.e. lists) were used, each participant received one of these six lists.

To ensure that participants track the events and infer the actual location of the kitten, twice during the practice and eight times during the test phase, they were asked to indicate the location of the kitten, after performing the rating (attention check trials). In specific, following the rating, they were presented with a central fixation cross, then a picture of three closed boxes with the question *'Where is the kitten?*' above and they had to click on the appropriate box as fast as possible. The picture of the boxes remained on screen for a maximum of 5000 ms or until a response was given. The presentation of these 'attention check' trials was counterbalanced and randomized, such that participants could not predict when they would have to indicate the location of the kitten, based on the perspective prompt, the belief type or the to-be-rated alternative. Except for the two practice trials, no feedback was

[21] Note that factor level labels refer to the modal status of the three types of alternatives, from different perspectives: actual – for the participant (on the UB and for both the participant and the agent on the TB trials); possible – for the agent (on the UB trials); impossible - for both the participant and the agent (on both types of trials). For reasons of brevity we simply use the term 'alternative' when referring to them in the text.

provided for them, but participants were warned that good performance on these trials is essential for their data to be accepted, before the test phase started.

4.2.1.5 Data analysis

Trials on which participants did not provide a response or failed to click on the fixation cross were considered invalid and, as such, were not included in the analyses (SELF: $M_{invalid}$ =1.28%, $SD_{invalid}$=2.94%; OTHER: $M_{invalid}$ =1.96%, $SD_{invalid}$=0.88%; for details see: Supplementary Materials **Table S1**). Responses for attention check trials were analysed only for hit rate, to check whether participants meet the inclusion criteria. Participants with <70% valid trials and those who provided incorrect or no response for more than 33% of the attention check trials (3 out of the 8) were not included in the analyses, as such results were considered to indicate lack of motivation or inattentiveness.

Responses were analysed in the following way: First, to be able to compare the ratings across participants and, at the same time, compensate for the individual differences in the scale use, we divided the x-coordinate of the participant's clicks by the participant's individual scale range (calculated by computing the difference of the maximum and minimum X-coordinate of the participant's clicks) to obtain the 'normalized relative position' of the cursor on the scale. Then, we averaged this measure, ranging from 0 to 1, per perspective, belief, separately for each type of alternative, across participants. Participants whose mean rating were <0.75 on the SELF perspective true belief actual or >0.25 on the SELF true belief impossible trials were excluded from the analyses, as such results were taken as signs of misunderstanding the task or the use of the scale. This resulted in the exclusion of N=1 participant. Given the non-normal distribution of the dependent variables and the extreme values (cumulating around 1 for the actual on the SELF and around 0 for the impossible alternative, on both the SELF and OTHER perspective trials) that made it impossible to normalize the data, ratings were analysed by appropriate nonparametric tests.

Our three crucial tests were: comparison of the ratings for the 1) actual and impossible alternative, on the SELF perspective true belief trials; 2) actual and possible alternative on the OTHER perspective underspecified belief trials; and 3) possible and impossible alternative on the SELF underspecified as well as the SELF true belief trials. While the first comparison tested whether participants themselves performed the necessary computations (i.e. inferred the location of the hidden animal), the second assessed the validity of the paradigm, i.e. whether participants were able to track what inferences the other might make on the basis of her knowledge, *explicitly*. Finally, the third comparison tested the actual hypothesis, whether adults also track others' inferences *spontaneously*. We hypothesized that if they do so, this should bias their ratings on the possible trials, from 'less' towards 'more likely'.

In addition to the group-level analyses we also computed the difference of mean ratings provided for the possible and impossible alternative, on the self underspecified and true belief trials, for each participant, to investigate the ratio of those in the sample who showed the predicted effect. If the difference of the two scores (SELF UB possible-impossible difference - SELF TB possible-impossible difference) was larger than 0.01, i.e. larger than 1% of the scale length, this was considered as a tentative evidence for spontaneously representing the alternatives (thereby the conclusion) of the other agent.

Reaction time (RT) data 2 standard deviations lower or above the given participant's mean RT for the respective experimental condition were excluded from the RT analyses (for trial exclusion rates see: Supplementary Materials **Table S4.3)**. Participants whose overall RT (averaged across all test trials) were 2 standard deviations lower or above the group mean were considered outliers and, as such, their RT data was not analysed. This resulted in the exclusion of N=2 participants. RTs were first log-transformed, and averaging was performed on the log-transformed RT data. Mean reaction times were investigated by running 2 x 3 repeated-measures ANOVAs, run separately for SELF and OTHER trials (applying Greenhause-Geisser correction whenever sphericity assumptions were not met), with belief and alternative type as within-subject factors, and subsequent paired-samples t-tests, focusing on the differences between participants' RTs on the actual versus possible trials in the OTHER and on the possible versus impossible trials in the SELF underspecified belief condition to capture potential egocentric and altercentric interference effects, respectively.

To adjust for multiple comparisons, both in the rating and the RT analyses, the Holm's Sequential Bonferroni Procedure was used (taking into account three conditions when performing the adjustments). Adjusted p-values were calculated in R (using the 'p.adjust' function of the stats package). All tests were two-tailed with significance level set at $p<0.05$.

4.2.2 Results

4.2.2.1 Rating analyses

Other-perspective trials

To test whether participants were able to track what conclusions the other agent might draw from the beliefs she holds when they track her inferences *explicitly,* first we analysed the ratings provided on the other-perspective trials. Friedman tests indicated a significant difference between the three types of alternatives, on both the true and the underspecified belief trials (TB: $\chi^2(2)=52.88$, $p< .001$, Kendall's $W=0.755$; UB: $\chi^2(2)=53.80$, $p< .001$, Kendall's $W=0.769$) (see **Figure 4.2**). Post hoc Wilcoxon Signed Rank tests revealed that on the true belief trials, when the agent saw the same events as participants did, participants' ratings were significantly higher for the actual than either for the possible ($Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$) or for the impossible alternative ($Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$), with the latter two rated similarly ($Z=-0.75$, $p_{adj(3)}=.451$, $r=0.127$), indicating that, in these cases, participants were aware of that in these cases the agent could arrive to the same conclusion they did (and correctly infer the location of the kitten). In contrast, on those other-perspective trials where the agent did not witness the second box opening, participants' ratings for the actual and the possible alternative did not differ significantly ($Z=-1.41$, $p_{adj(3)}=.159$, $r=0.238$), although the mean rating was somewhat higher for the actual than for the possible alternative. Participants rated the likelihood of these two alternatives significantly higher than the likelihood of the impossible alternative (actual-impossible: $Z=-5.16$, $p_{adj(3)}<.001$, $r=0.872$; possible-impossible: $Z=-5.16$, $p_{adj(3)}= .001$, $r=0.872$), with the mean ratings cumulating around 0.5-0.6 for both alternatives, suggesting that they understood that these two were equally likely for the agent.

Self-perspective trials

To check whether participants themselves performed the necessary computations (i.e. the disjunctive inference from first-person perspective) and to test for our main hypothesis (whether human adults also track others' inferences *spontaneously*), next, we analysed the likelihood estimations provided on the self-perspective trials.

Friedman tests indicated a significant overall difference between the three types of alternatives on both the true ($\chi^2(2)=53.28$, $p< .001$, Kendall's $W=0.761$) and the underspecified belief trials ($\chi^2(2)=55.1$, $p< .001$, Kendall's $W=0.796$). As can be seen on **Figure 4.2**, ratings were significantly higher for the actual than either for the possible or for the impossible alternative, on both types of belief trials (TB – actual-possible: $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; actual-impossible: $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; UB – actual-possible: $Z=-5.23$, $p_{adj(3)}< .001$, $r=0.872$; actual-impossible: $Z=-5.23$, $p_{adj(3)}< .001$, $r=0.872$), indicating that participants could safely infer the kitten's actual location. Crucially, however, on trials where the agent did not witness the second opening, hence could eliminate only one alternative, participants also provided higher ratings for the possible compared to the impossible alternative, providing evidence for an altercentric bias. Although this shift in participants' estimations from less towards more likely on the scale was quite small ($M_{possible}=0.06$, $SD_{possible}=0.13$ versus $M_{impossible}=0.02$, $SD_{impossible}=0.03$) the difference between the ratings provided for the two types of alternatives was significant ($Z=-2.46$, $p_{adj(3)}= .014$, $r=0.415$). No such difference was present on the true belief trials ($Z=-0.26$, $p_{adj(3)}= .796$, $r=0.044$), suggesting that the effect was indeed the result of the mismatch between the two conclusions (the one made from first-person perspective and the one spontaneously computed for the other). Inspection of the individual means revealed that the predicted effect was present in N=8 participants (SELF UB possible-impossible difference minus SELF TB possible-impossible difference > 0.01, i.e. 1% of the scale length). Using a more lenient measure (SELF UB impossible-possible difference score minus SELF TB impossible-possible difference score > 0), altogether N=20 participants demonstrated a rating pattern that was in line with our hypotheses.

The results remained the same after removing one participant from the analysis who could be considered an outlier based on his/her mean rating on the SELF UB possible trials (> 0.50; SELF true belief trials – actual-impossible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; possible-impossible: $Z=-0.39$, $p_{adj(3)}= .695$, $r=0.067$; SELF underspecified belief trials - possible-impossible: $Z=-2.27$, $p_{adj(3)}= .023$, $r=0.389$; OTHER underspecified belief trials – possible-impossible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; actual-possible: $Z=-1.41$, $p_{adj(3)}= .158$, $r=0.242$). Excluding those participants who could be considered outliers on the basis of their average RTs (N=2), and thus were not included in the RT analyses, did not change the main results either (SELF true belief trials – actual-impossible: $Z=-5.01$, $p_{adj(3)}< .001$, $r=0.873$; possible-impossible: $Z=-0.16$, $p_{adj(3)}= .875$, $r=0.028$; SELF underspecified belief trials - possible-impossible: $Z=-2.78$, $p_{adj(3)}= .005$, $r=0.484$; OTHER underspecified belief trials – possible-impossible: $Z=-5.01$, $p_{adj(3)}< .001$, $r=0.873$), although a marginally significant difference emerged between the actual and the possible alternative on the other-perspective underspecified belief trials ($Z=-1.71$, $p_{adj(3)}= .088$, $r=0.298$).

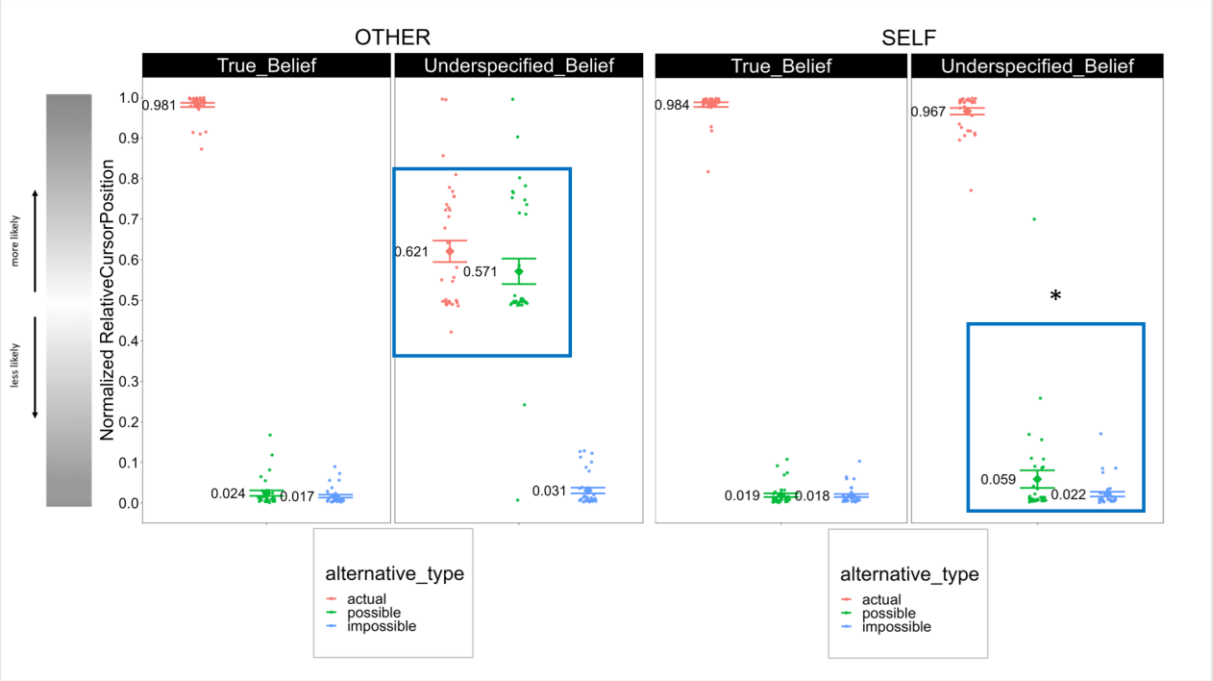


Figure 4.2 Mean ratings (mean normalized relative cursor position) on the self- and other perspective trials per belief and alternative type in Experiment 1. Higher values indicate that participants considered it more likely that the kitten was hiding at that specific location, either from their own (SELF trials) or from the agent's perspective (OTHER trials). Error bars represent 95% CI, dots show the individual means. Blue frames indicate our two main foci of interest: the difference between the possible and impossible alternative on the self-perspective and the actual and possible alternative on the other-perspective underspecified belief trials. The first comparison investigates whether there is an altercentric bias in the estimations (indicating spontaneous representation of what the other considers possible) on the self-perspective trials, and the second test for the explicit representation of the two alternatives the agent could represent on the other-perspective trials. Stars indicate significant differences between the two experimental conditions. *: $p<0.05$, **: $p<0.01$

4.2.2.2 Reaction time analyses

To capture the potential impact of participants' own conclusion on how easy they perform the judgements from the other's perspective, in case the two does not match, and that of the other agent on their own decision making process (i.e. potential egocentric and altercentric interference effects), next, we performed a 2 x 3 repeated-measures ANOVA with belief (true versus underspecified) and alternative type (actual, possible, impossible) as within-subject factors, first on the other then, separately, on the self-perspective trials.

Other-perspective trials

Analysis of the OTHER trials yielded a significant main effect of belief ($F(1, 32)=16.5$, $p< .001$, $\eta_p^2= .338$), resulting from the fact that participants were generally slower to provide their ratings on the underspecified than on the true belief trials, but no significant main effect of alternative type ($F(2, 64)=1.42$, $p= .250$, $\eta_p^2= .042$) or belief x alternative type interaction ($F(2,64)=0.67$, $p= .935$, $\eta_p^2= .002$), reflecting the fact that it took roughly equal time for participants to rate the three types of alternatives, both on the true and the underspecified belief trials. In line with this, none of the pairwise comparisons were significant (TB: all *t*s<0.83, all p_{unadj}s> .416; UB all *t*s<1.33, all p_{unadj}s > .192). These results indicate that, although it took some effort to perform the judgements from the other agent's different perspective, participants had no difficulties with tracking her logical inferences, in an explicit manner, and representing what conclusions she could (and could not) draw from the evidence she had.

Self-perspective trials

Analysis of the self-perspective trials yielded somewhat different results: no significant main effect of belief ($F(1 ,32)=0.40$, $p= .530$, $\eta_p^2= .012$) or alternative type ($F(1.55,49.48)=2.33$, $p= .120$, $\eta_p^2= .068$). There was no significant belief x alternative type interaction either ($F(2, 64)=0.69$, $p= .507$, $\eta_p^2= .021$). As can be seen on the graph, it took longer for participants to estimate the likelihood of the kitten being at the 'possible' location, i.e. in the box that opened second (*M*=2081 ms, *SD*=306 ms) than to perform their ratings for the impossible location, i.e. for the box that opened first (*M*=1977 ms, *SD*=257 ms) on the UB trials, i.e. when the other agent did not witness the second box opening (therefore could consider the second box a potential hiding place for the kitten), suggesting altercentric interference from the other's perspective. After correcting for multiple comparisons, the difference was, however, not significant ($t(32)=2.06$, $p_{\text{unadj}}= .048$, $p_{\text{adj(3)}}= .144$, $d=0.358$; for the other two comparisons: *t*s< 1.52, p_{unadj}s >.140). On the TB trials, pairwise comparisons revealed no significant difference between the three types of alternatives with respect to how long it took participants to rate those (all *t*s<0.78, all p_{unadj}s> .442). There was no significant difference between the two types of beliefs along the mean RTs provided for any of the three alternatives (all *t*s<1.06, all p_{unadj}s> .295).

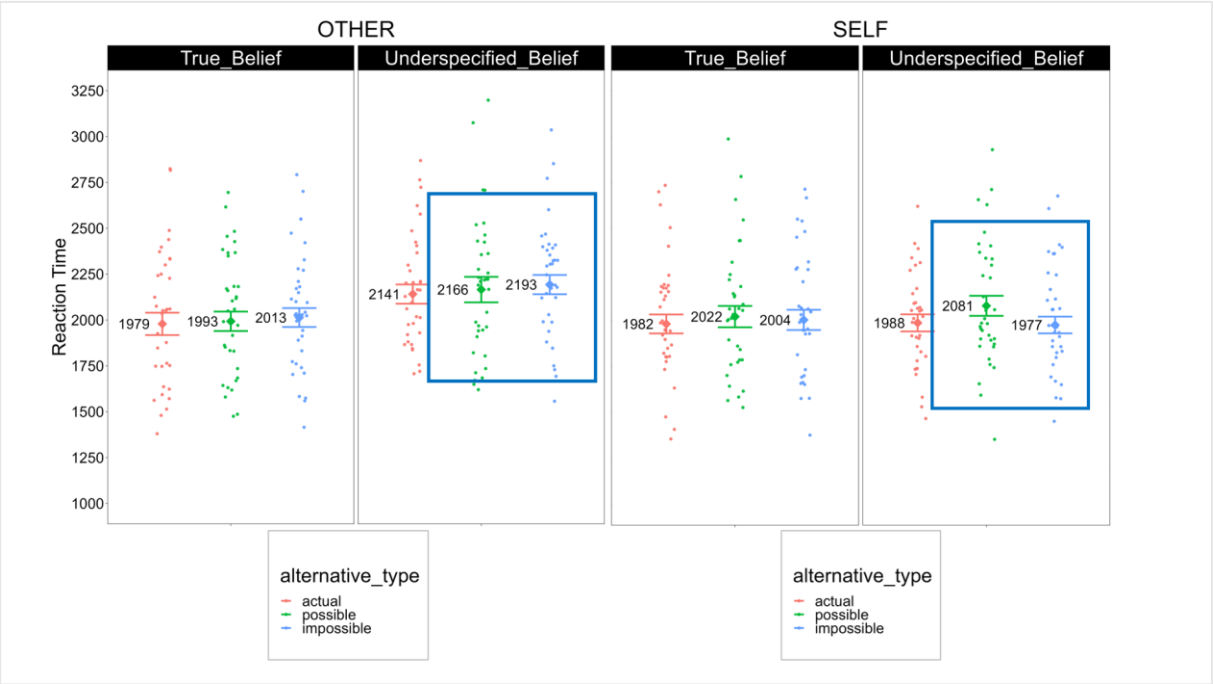


Figure 4.3. Mean reaction times (time necessary to perform the likelihood estimations) on the self- and other perspective trials per belief and alternative in Experiment 1. Error bars represent 95% CI, dotsand show the individual means. Blue frames indicate our two main foci of interest: the difference between mean rating time of the possible and impossible alternative on the self- and on the other-perspective underspecified belief trials.

4.2.3 Discussion

Results observed on the 'OTHER' trials, on which participants had to track other agents' beliefs in an explicit manner, indicate that human adults can and do take into account what conclusions other agents may draw from the information they have access to (and the belief they form based on that information). On trials where they had to judge the likelihood that the kitten has hidden at the presented location from the other agent's perspective and the agent did not see the second box opening, participants' ratings were similar for the box that opened second and for the box that remained closed at the end, i.e. for the two locations the agent could consider a potential hiding place for the kitten, with their likelihood estimations clearly indicating that they attributed a belief with an underspecified belief content to her (a disjunction: '*the kitten is either in Box A or Box B'*).

Crucially for our hypothesis, participants did not only assign a higher probability to the animal's occurrence at the possible location (compared to the impossible one) when they had to take the other agent's perspective, but also when this was not necessary, i.e. on the self-perspective trials, when the agent could represent two, equally likely alternatives, suggesting that they *spontaneously* represented what conclusions the other agent could draw in the given situation, on the basis of what events she had witnessed. The pattern of reaction times pointed in the same direction, indicating an altercentric intrusion of the other agent's belief content (i.e. longer RTs), on those self-perspective trials where the agent was uncertain about the location of the animal, although the difference between the two critical conditions did not remain significant after correcting for multiple comparisons. Taken together these results indicate that human adults may not only encode what another agent sees, believes or knows, in a spontaneous manner, but also what the other agent might infer from the represented content, at least in situations where the other agent lacks a certain piece of information that would be necessary to make the appropriate inference and to arrive to the same conclusion the participant does.

Nevertheless, it is important to point out that the effect was rather small and the group-level difference was mainly driven by the ratings of 8 participants, for whom the 'altercentric bias' was more pronounced than for the others (though another 12 also demonstrated a rating pattern that was in line with our hypothesis). Therefore, we decided to run a conceptual replication of Experiment 1. Specifically, we tested whether the altercentric bias observed in Experiment 1 emerges if the information participants have to track and infer is (hence the disjunction they have to attribute to the other agent on the underspecified belief trials is about) an object's identity.

4.3. Experiment 2

Experiment 2 differed from Experiment 1 in the following aspects: 1) three different animals 'hid' in the boxes (equally often at each location) and the opening of the two boxes revealed which animal is hiding inside; 2) participants had to infer which of the three animals is hiding in the third box that remained closed, after seeing which animal is hiding in Box A and Box B; 3) and they had to rate the likelihood that a certain animal (presented on a picture) has hidden in the scene, at the end. That is, it differed from Experiment 1 both in the type of information participants had to track and infer (the identity of the hidden animal) and in the type of information they received (about the presence instead of the absence of animals at certain locations). Consequently, the inferential process participants had

to go through in Experiment 2, involved an additional inferential step compared to Experiment 1. Specifically, unlike in Experiment 1, after representing all three alternatives (step 1: '*Box C contains either the fox, the chick or the frog'*), in Experiment 2 participants had to first consider the evidence they received about the first two boxes (step2: '*Box A contains the fox, Box B contains the chick'*), and make an inference from this evidence, to be able to eliminate two alternatives as candidate 'contents' of the closed box (step3: '*therefore the content of Box C is neither the chick nor the fox*'), and draw the final conclusion (step4: '*therefore it is the frog'*). In Experiment 1, they could directly proceed to step 3 and step 4, skipping step 2, due to the form of the evidence they received (negative versus positive)[22], which made the inferential process somewhat longer / less complex. Apart from these differences, the general structure of the task was the same as the one used in the first experiment as well as the general inferential rule participants had to apply (disjunctive reasoning), to perform the inference from self-perspective and to attribute the appropriate belief content to the other. We hypothesized that if participants track other agents' inferences not only about object location but also about object identity, they should provide higher ratings for the animals presented in the second box (compared to the ones presented in the first box) and/or should have longer RTs, on the underspecified belief trials, i.e. when the agent did not witness the second-box opening.

4.3.1 Methods

4.3.1.1 Participants

The final sample consisted of 34 participants (M_{age}=26.91, SD_{age}=4.96, 18 males). Participants were recruited via Testable Minds. All of them had at least high school degree, 5 were left-handed, 2 ambidextrous, the rest was right-handed. The inclusion criteria were identical to the ones used in Experiment 1. Ten more participants were tested but were not included in the analyses: 2 were excluded because their rating on the self- perspective TB actual and/or the self- perspective TB impossible trials indicated lack of understanding of the task or the use of the rating scale, further 8

[22] Note that, in theory, it is possible that participants in Experiment 1 represented that *'Box A and Box B are empty*' and made an inference on the basis of the evidence presented (*'therefore the kitten is neither in Box A nor in Box B*'), just like in Experiment 2. However, this is not how inferences are traditionally assumed to take place in reasoning by exclusion tasks.

participants were not included in the analyses, either because they had less than 70% valid trials (N=3), or because they failed to answer >50% of the attention check trials correctly (N=5). The study was approved by the EPKEB United Ethical Committee, Hungary; participants received monetary compensation of 5.3 USD for their participation. Experiment 2 was preregistered on AsPredicted.org (with the document number 83897); the sample size was determined to match that of Experiment 1.

4.3.1.2 Stimuli and apparatus

The images used in the training session and on the experimental trials of the main task were similar to the ones used in Experiment 1, with two key differences. First, on those images where one or two of the boxes were open, the open boxes contained one or two of three animals (a red fox, a green frog or a yellow chick), respectively, with only their head protruding from the box(es) (see **Figure 4.4a** for examples). The different location and animal (or animal-pair) combinations resulted in three versions of each one-box open and six versions of each two-box open picture. Second, all three boxes had the same colour, to prevent participants from encoding the box's colour, in addition to the identity of the animal, and generally could have diverted focus from the relevant information participants had to track. The stimuli used on the attention check trials included the images of the three animals, the frog, the fox and the chick used in the main task (each extending 160 x 200 pixel in height and width), arranged horizontally, at equal distance from each other.

Active screen area sizes ranged from 1280x720 pixels to 2560 x 1440 pixels (data collected by Testable).

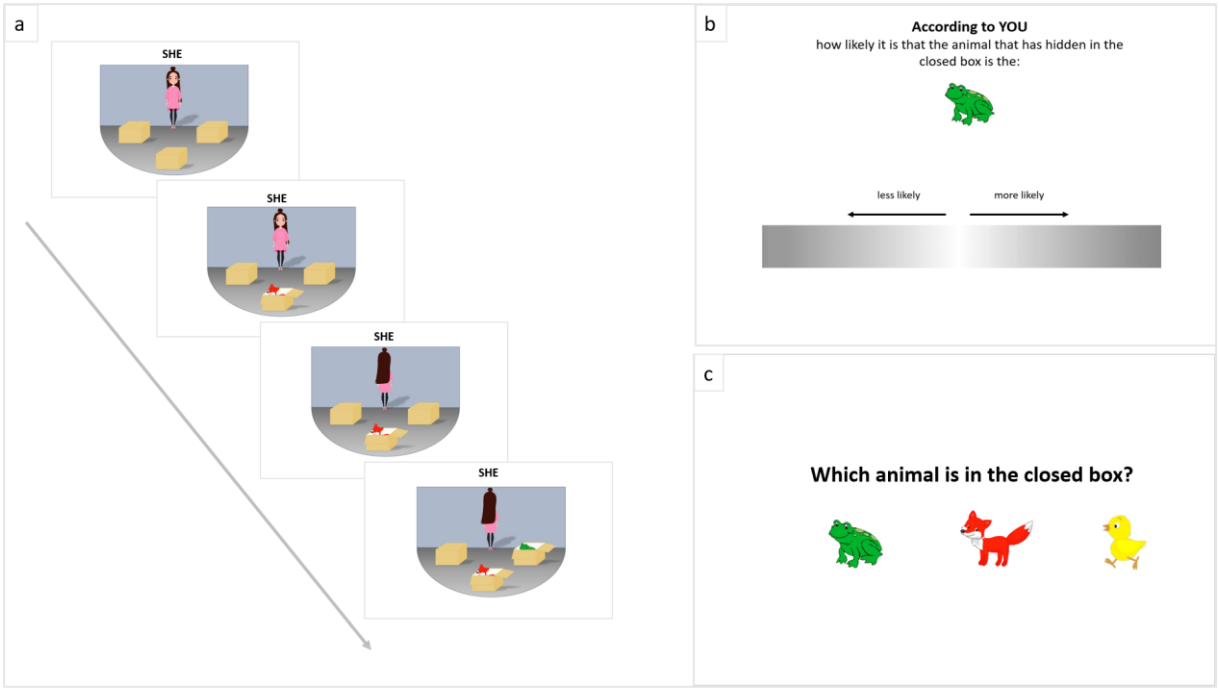


Figure 4.4. (a) An example of a picture sequence presented in Experiment 2. The trial structure was the same as in Experiment 1. Trials started with a fixation cross which was followed by a perspective prompt (the word YOU or SHE). After the presentation of the fourth picture, participants were presented with the rating scale along with the question and the picture of the to-be-rated animal above. (b) An example of the response screen in the main task's experimental trials. (c) The response screen of attention check trials in Experiment 2.

4.3.1.3 Procedure

The general structure of the task was the same as the one used in Experiment 1 with one exception: as the two boxes that opened always contained an animal, there was no need to train participants to recognize when they are empty, hence participants received only one training session (the second training session from Experiment 1), that aimed to ensure that participants understand that when the agent is not witnessing the second box opening, she is uncertain about which of the two remaining animals hide in the second-open and the third, closed box. Just like before, each phase was preceded by its own instruction screen (see Supplementary Materials **S4.1.2**), with a maximum time limit set, to prevent participants from disengaging from the task for longer time periods.

Training session

At the beginning of the session, participants were told that a frog, a fox, and a chick has hidden in the scene (showing their images) and, at the end of each trial, they would have to decide whether the girl

would search for the specific animal named in the question, in one or two boxes. The structure of the training trials was the same as before, with the only difference being that the question at the end referred to a specific animal (i.e. *'Would the girl search for the FOX in*: ONE or TWO boxes'). Participants received altogether six underspecified belief and three true belief trials, in a fixed pseudorandom order, with each animal appearing (and being hidden) at each location (i.e. in the left, the right and the central box) three times, to ensure that participants do not form associations between the location of the boxes and the identities of the animals. In case they failed to respond correctly on more than two underspecified belief trials or on more than one true belief trial, the session was repeated once.

Main Task

The trial structure of the main task was the same as in Experiment 1, with one difference, resulting from the fact that this was an identity and not a location task. At the end of the trials, participants had to rate how likely is that a certain animal has hidden in the scene and, instead of the picture of the scene, with the closed boxes, the target question (*'According to YOU/the GIRL how likely it is that the animal that has hidden in the closed box is the'*) was presented together with the picture of one of the three animals (see **Figure 4.4b**). In addition, to decrease the number of excluded trials, participants were required less precision when clicking at the fixation cross (we told them simply that they have to click on the fixation cross and not *exactly at the middle of the* cross, as we did in the first experiment). Other aspects of the design were the same as before, with alternative types defined by whether the question was about the animal hiding in the closed box (actual), the animal that was presented first (impossible), or the animal that was presented second (possible). The perspective, the belief type, the alternative type, the identity and the location of the to-be-rated animal were counterbalanced within block as well as the identity and the location of the animal that was actually hidden, that was first presented, as well as the location of the three different animals, such that each animal was presented equal times at each of the three locations. The order of trials was pseudorandomized such that there were no more than three consecutive trials with the same perspective, belief, alternative type, animal question and picture sequence starting with the presentation of the same animal in a row. Eight different trial order was used, the presentation of which was counterbalanced across participants.
As in Experiment 1, in addition to the experimental trials, participants received 8 attention check trials, to ensure that they perform the inference from first-person perspective. In specific, following a central fixation cross, they were presented with the pictures of the three animals, with the question '*Which animal is in the closed box?*' above, and were asked to indicate the identity of the hidden animal by

clicking on the image of one of the animals, as fast as possible. The counterbalancing and the pseudorandomization of the attention check trials were the same as in the first experiment.

4.3.1.4 Data analysis

The data was analysed as before, using the same data and participant exclusion criteria as in Experiment 1[23]. In addition, we also compared the two experiments along the (1) difference scores computed from the mean ratings of (a) the possible and impossible trials of the self-perspective underspecified belief and (b) the possible and impossible as well as the actual and possible trials of the other-perspective underspecified belief trials on the one hand, and 2) along the mean reaction times necessary to perform the ratings on the other, to investigate whether the change in the information that had to be tracked and the type of belief content that had to be computed, had an effect on the tracking of the other agent's inferences. The difference scores were compared by running a series of Mann-Whitney tests. Potential differences in the mean RTs were investigated by running mixed ANOVAs on the log-transformed RT data, with experiment as a between-subject and belief and alternative type as within-subject factors, separately for the self- and other-perspective trials.

[23] The data analysis deviated from the preregistration in two ways. First, the preregistration did not include exclusion criteria, only the two inclusion criteria for participants. Second, in the preregistration we defined the relative cursor position (our main dependent measure) as the mean of the X-coordinate of the clicks/the participant's screen width. Later, however, we realized that we i) also need a participant exclusion criteria, given that some participants simply did not follow the instructions (due to misunderstanding the task, the use of the scale or due to not paying attention to the events) and that it is better to ii) use the normalized relative cursor position as our measure (given that the absence of clear endpoints on the scale allows for larger than usual variations in how the different individuals use the scale, and one should compensate for this to be able capture potentially subtle differences).

4.3.2 Results

4.3.2.1 Rating analyses

Other-perspective trials

Regarding the OTHER trials, which investigated whether participants represent what conclusions another agent may draw regarding an object's identity, when tracking her beliefs deliberately, Friedman tests indicated a significant difference between the three types of alternatives, on both the true and underspecified belief trials (TB: $\chi^2(2)=53.17$, $p< .001$, Kendall's $W=0.782$; UB: ($\chi^2(2)=53.88$, $p< .001$, Kendall's $W=0.792$). As in Experiment 1, on the true belief trials, participants' ratings were significantly higher for the actual than either for the impossible ($Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$) or for the possible alternative ($Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$), with the means cumulating around 1 for the actual and 0 for the other two alternatives, indicating that participants understood that when the agent witnessed the same events as they did, she could arrive to the same conclusion regarding the identity of the hidden animal (see **Figure 4.5**). Interestingly, on these trials, participants also considered it somewhat more likely that, according to the agent, the hidden animal was the one that was in the box that opened second than that it was the one that was in the box that opened first ($Z=-1.92$, $p_{adj(3)}= .055$, $r=0.29$), although the difference was very small and may just reflect a carry-over effect from the UB trials, where this animal was a possible alternative for the other.

On the underspecified belief trials, when the agent did not witness the second opening, participants' ratings were similar for the animals hiding in the second and in the third box ($M_{possible}=0.59$, $SD_{possible}=0.17$ versus $M_{actual}=0.65$, $SD_{actual}=0.17$), suggesting that they understood that, in this situation, the agent represented two, roughly equally likely, alternatives. In specific, participants' ratings were significantly higher for both the actual and the possible than for the impossible alternative, on these trials (actual-impossible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; possible-impossible: $Z=-2.98$, $p_{adj(3)}= .003$, $r=0.290$), with the individual means cumulating around 0.6 for both alternatives. Nevertheless, this does not mean that participants rated the two alternatives the same way. Despite the actual and the possible alternative was equally likely for the other agent, participants provided somewhat higher ratings for the former than for the latter, i.e. for the animal that was actually hiding in the third box than the animal that was hiding in the box the opening of which was not witnessed by the agent ($Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$), suggesting a difficulty with inhibiting their own knowledge.

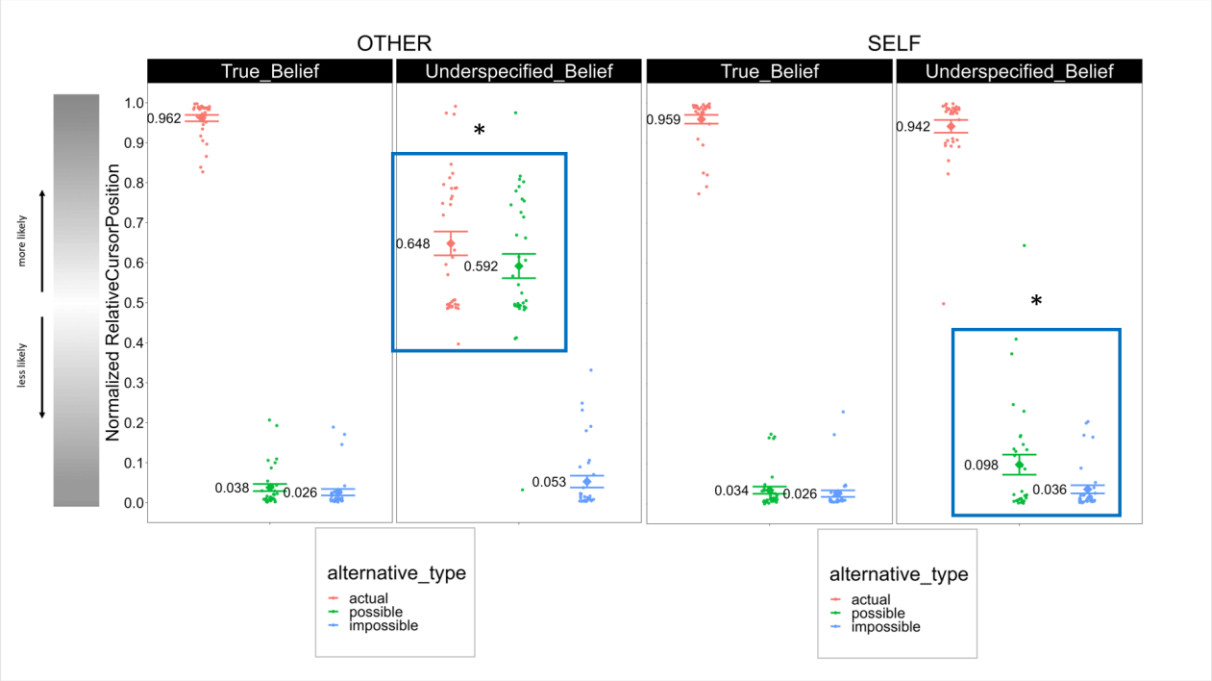


Figure 4.5 Mean ratings (mean normalized relative cursor position) on the self- and other perspective trials per belief and alternative type in Experiment 2. Higher values indicate that participants considered it more likely that the presented animal was hiding in the third, closed box, either from their own (SELF trials) or from the agent's perspective (OTHER trials). Error bars represent 95% CI, dots show the individual means. Blue frames indicate our two main foci of interest: the difference between the possible and impossible alternative on the self-perspective and the actual and possible alternative on the other-perspective underspecified belief trials. The first comparison investigates whether there is an altercentric bias in the estimations (indicating spontaneous representation of what the other considers possible) on the self-perspective trials, and the second test for the explicit representation of the two alternatives the agent could represent on the other-perspective trials. *: p<0.05, **: p<0.01

Self-perspective trials

As for the self-perspective trials, which addressed our main research question, specifically whether adults track other agents' inferences regarding object identity spontaneously, analyses indicated a significant overall difference between the three types of alternative on both trialtypes, just like in Experiment 1 (TB: $\chi^2(2)=52.10$, $p< .001$, Kendall's $W=0.766$; UB: $\chi^2(2)=51.53$, $p< .001$, Kendall's $W=0.758$). Ratings were significantly higher for the actual than for the other two types of alternative, on both the true (actual-possible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; actual-impossible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$) and the underspecified belief trials (actual-possible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; actual-impossible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$), indicating that participants had no problems with inferring the hidden animal's identity.

Crucially, as in Experiment 1, on trials where the agent did not witness the opening of the second box, participants rated it more likely that the hidden animal was the one hiding in there than that it was the one hiding in the first box (the opening of which was witnessed by the agent). Although the difference between the two conditions was relatively small ($M_{possible}$=0.10, $SD_{possible}$=0.14 versus $M_{impossible}$=0.04, $SD_{impossible}$=0.06), it was significant (Z=-1.98, $p_{adj(3)}$= .048, r=0.340). No such difference was present on the true belief trials (Z=-1.40, $p_{adj(3)}$= .161, r=0.240). Inspection of the individual data revealed that the predicted effect was present in N=14 participants, according to the stricter criteria, described in Section 4.2.1.5 of Experiment 1 (with altogether N=18 participants demonstrating a difference score on the self-perspective UB and TB trials that was in line with our predictions, according to the leaner criteria).

After excluding one participant, who could be considered an outlier based on his/her mean rating on the SELF underspecified belief possible trials (> 0.50), the difference between ratings provided on the SELF UB possible and impossible trials became only marginally significant ($M_{possible}$= 0.081, $SD_{possible}$=0.019 versus $M_{impossible}$= 0.07, $SD_{impossible}$=0.010; Z=-1.76, $p_{adj(3)}$= .078, r=0.307). There was no change in the rest of the results (SELF true belief trials – actual-impossible: Z=-5.01, $p_{adj(3)}$< .001, r=0.892; possible-impossible: Z=-1.18., $p_{adj(3)}$= .239, r=0.205; OTHER underspecified belief trials – actual-possible: Z=-3.05, $p_{adj(3)}$= .002, r=0.531; possible-impossible: Z=-5.01, $p_{adj(3)}$< .001, r=0.892). Excluding those participants who were not included in the RT analyses below (N=3), did not change the main results (SELF true belief trials – actual-impossible: Z=-4.85, $p_{adj(3)}$< .001, r=0.856; SELF underspecified belief trials - possible-impossible: Z=-2.25, $p_{adj(3)}$= .024, r=0.396; OTHER underspecified belief trials – actual-possible: Z=-4.86, $p_{adj(3)}$< .001, r=0.856; possible-impossible: Z=-4.86, $p_{adj(3)}$< .001, r=0.856). Although a tendency level difference emerged between the possible and the impossible trials also in the SELF true belief condition (Z=-1.66., $p_{adj(3)}$= .098, r=0.292), its magnitude was smaller than the magnitude of the effect on the SELF underspecified belief trials (UB difference score vs TB difference score: Z=-1.72, p_{unadj}= .085, r=0.303).

4.2.2.2 Reaction time analyses

Other-perspective trials

To test whether participants' own conclusions interfere with how easy they perform their judgements from the agent's perspective, next we analysed participants rating times of the other-perspective

trials. A 2 x 3 repeated measures ANOVA with belief (TB versus UB) and alternative type (actual, possible, impossible) as within-subject factors, revealed a significant main effect of alternative type ($F(2, 60)=8.56$, $p= .001$, $\eta_p^2= .222$), resulting from the fact that it took longer for participants to perform their estimations for the impossible than for the other two types of alternative (i.e. for the animal hiding in the first box that opened), on both the true and the underspecified belief trials (see **Figure 4.6**). There was, however, no significant main effect of belief ($F(1, 30)=1.69$, $p= .204$, $\eta_p^2= .053$), despite the generally longer RTs on the underspecified belief trials. There was no significant belief x alternative type interaction either ($F(2, 60)=0.64$, $p= .529$, $\eta_p^2= .021$), reflecting the similar reaction time pattern on the two types of belief trials. In line with this, pairwise comparisons indicated marginally significant difference between the impossible and the other two types of alternative on the TB (actual-impossible: $t(30)=2.24$, $p_{adj(3)}= .066$, $d=0.402$; possible-impossible: $t(30)=2.44$, $p_{adj(3)}= .063$, $d=0.438$) and significant difference on the UB trials (actual-impossible: $t(30)=3.02$, $p_{adj(3)}= .005$, $d=0.541$; possible-impossible: $t(30)=2.38$, $p_{adj(3)}= .048$, $d=0.427$), but no significant difference between the time necessary to perform the ratings for the actual and the possible alternative, on either of the two trialtypes (TB: $t(30)=0.42$, $p_{adj(3)}= .675$, $d=0.076$; UB: $t(30)=-0.53$, $p_{adj(3)}= .598$, $d=0.096$). This suggests that participants had no difficulty in tracking the other's logical inferences and monitoring what she considers possible in a given situation, *explicitly*, when taking her perspective.

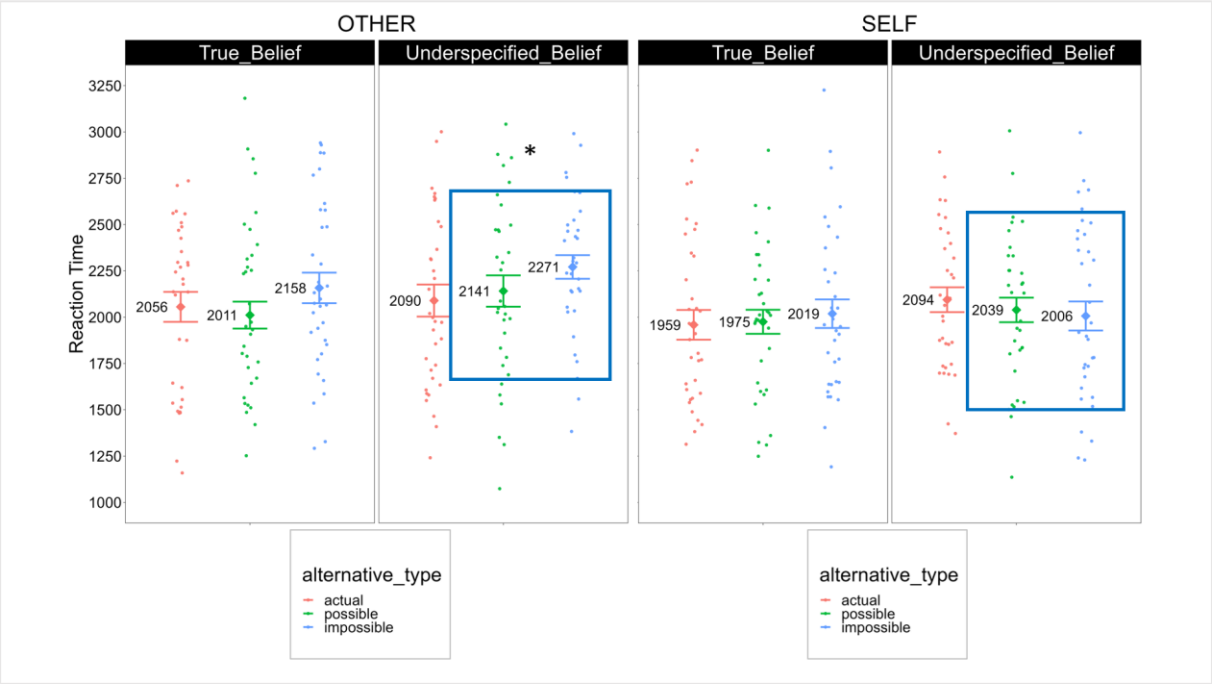


Figure 4.6. Mean reaction times (time necessary to perform the likelihood estimations) on the self- and other perspective trials per belief and alternative type in Experiment 2. Error bars represent 95% CI, dots show the individual means. Blue frames indicate our two main foci of interest: the difference between mean rating time of the possible and impossible alternative on the self- and on the other-perspective underspecified belief trials.

Self-perspective trials

Finally, to investigate our main hypothesis, whether the agent's conclusion interferes with participants' own decision-making process, in case there is a mismatch, i.e. whether there are further signs of spontaneous tracking of the other agent's inferences, we performed the same analyses as above, on the self-perspective trials. A 2 x 3 repeated measures ANOVA with belief (TB versus UB) and alternative type (actual, possible, impossible) yielded no significant main effect of alternative type ($F(2, 60)=0.003$, $p= .997$, $\eta_p^2= .000$) but a tendency level main effect of belief ($F(1, 30)=3.69$, $p= .064$, $\eta_p^2= .109$) and belief x alternative type interaction ($F(1.67, 49.50)=2.83$, $p= .078$, $\eta_p^2= .086$). As can be seen on the graph, participants were somewhat slower to provide their ratings for the actual and the possible alternative when the agent was uncertain about the hidden animal's identity compared to the situation when she could infer which animal was hiding in the third box (actual – M_{UB}=2094 ms, SD_{UB}=365 ms versus M_{TB}=1959 ms, SD_{TB}=438 ms; possible – M_{UB}=2039 ms SD_{UB}=361 ms versus M_{TB}=1975 ms, SD_{TB}=354 ms), indicating some kind of an interference from the other agent's perspective. Pairwise comparisons revealed a significant difference between the TB and UB trials in how fast participants performed their estimations for the actual ($t(30)=-3.06$, $p_{adj(3)}= .015$, $d=0.550$) but not in how fast they rated the other two types of alternative (possible: $t(30)=-0.92$, $p_{adj(3)}= .732$, $d=0.165$; impossible: $t(30)=0.33$, $p_{adj(3)}= .72$, $d=0.060$). Crucially, the reaction time pattern was similar on the true and underspecified belief trials, with follow-up t-tests indicating no significant difference between the three types of alternatives on either of the two trialtypes (TB: all ts<1.49, all p_{unadj}s> .147; UB: all ts<1.45, all p_{unadj}s> .158). Despite the absence of the expected altercentric interference effect (significant difference between the rating times of the possible and the impossible alternative on the UB trials), altogether these results indicate spontaneous consideration of the other's perspective, at least the fact that her knowledge state differs from that of the participants. Hence, they corroborate the findings in our rating measure.

Comparison of Experiment 1 and Experiment 2

Rating analyses

Pairwise comparisons indicated no difference between the two experiments in terms of the magnitude of the predicted effect on either the SELF or the OTHER underspecified belief trials (Exp1 vs Exp – SELF UB possible-impossible difference score: U=562.00, p_{unadj}= .692, r=0.048; OTHER UB possible-impossible difference score: U=587.00, p_{unadj}= .924, r=0.012; OTHER UB actual-possible difference score: U=485.00, p_{unadj}=.187, r=0.159). Although the number of participants demonstrating the effect (according to the stricter criteria), was somewhat higher in Experiment 2 than in Experiment 1 (N_{Exp2}=14 versus N_{Exp1}=8, $\chi^2(2, N=69)=2.67$, p= .010, Cramer's V=0.197), there was no difference between the two experiments in the number of those whose difference score was in line with our predictions (N_{Exp2}=18 versus N_{Exp1}=20, $\chi^2(2, N=69)=0.12$, p= .726, Cramer's V=0.42), i.e. when using the leaner criteria, in line with the group-level findings that the two experiments did not differ in the extent of the altercentric effect.

Reaction time analyses

Analysis of the other-perspective trials yielded a significant main effect of belief ($F(1, 62)=10.98$, p= .002, η_p^2= .150) and alternative type ($F(2, 124)=9.55$, p< .001, η_p^2= .019), but no significant belief x alternative type interaction ($F(2, 124)=0.52$, p= .594, η_p^2= .008). There was no significant main effect of experiment ($F(1, 68)=0.004$, p= .953, η_p^2= .000), experiment x belief ($F(1, 62)=.278$, p= .28 η_p^2= .019) or experiment x belief x alternative type interaction either ($F(2, 124)=0.32$, p= .729, η_p^2= .005). There was, however, a significant experiment x alternative type interaction ($F(2, 124)=3.15$, p= .046, η_p^2= .048), resulting from the fact that it took somewhat longer for participants to perform their estimations on the impossible trials (i.e. for the alternative that could be excluded first) from the agent's perspective in Experiment 2, than in Experiment 1, independent of the belief of the other agent. Follow-up pairwise comparisons revealed that the difference between the two experiments was, however, not significant (impossible trials – TB: Welch's $t(53.38)=-1.03$, p_{unadj}= .308, d=0.259; UB: $t(62)=-0.72$, $p_{unad)}$= .473, d=0.154).

Analysis of the SELF trials revealed a tendency level main effect of belief ($F(1, 62)=3.67$ p= .060, η_p^2= .056), no significant main effect of alternative type ($F(1.79, 111.06)=0.71$, p= .482, η_p^2= .011), and a tendency level belief x alternative type interaction ($F(2, 124)=5.12$, p= .097, η_p^2= .027). Despite the

slightly different effect of the agent's underspecified belief content on participant's reaction times in the two experiments, there was no significant experiment x alternative type (F(1.79, 111.06)=0.7, p= .464, η_p^2= .012), experiment x belief (F(1, 62)=1.27, p= .265, η_p^2= .020) or experiment x belief x alternative type interaction (F(1.95, 121.15)=1.46, p= .236, η_p^2= .023). There was no significant main effect of experiment either (F(1, 62)=0.13, p= .722, η_p^2= .002): even though participants had to perform slightly more complex inferences in Experiment 2 than in Experiment 1, their reaction times did not differ markedly from the ones observed on the self-trials of the first experiment.

Taken together, these results suggest that tracking others' logical inferences about object identity may be of comparable ease as monitoring their logical inferences about the location of objects, whether these computations are performed explicitly or in a spontaneous manner.

4.3.3. Discussion

Experiment 2 replicated and extended the findings of Experiment 1 by providing evidence that human adults may not only track other agents' inferences regarding the location but also about the identity of objects, spontaneously, and may do so even if they have to apply a somewhat different type of disjunctive inference (which requires them to perform an additional inferential step prior drawing the conclusion from the premises). In specific, it demonstrated a similar 'altercentric' bias in participants' likelihood estimations, on the self-perspective trials, to what we found in Experiment 1, in a situation where participants had to infer the identity (instead of the location) of a hidden object, from the information they received about the presence (rather than the absence) of certain objects at specific locations in the scene. Participants provided higher likelihood ratings for the alternatives they could exclude from the range of options, but the agent considered 'possible' (compared to the ratings of those alternatives both of them could eliminate from the range of options), not only on those trials where they had to track the agent's belief but also when this was not necessary, as they were performing ratings from self-perspective. Though the individual variability was very high (and hence, in some analyses, the difference between the two critical conditions was only marginally significant), the mean rating of the possible alternative, reflecting the altercentric intrusion of the other's perspective, was even higher in this than in the previous experiment.

In line with these findings, reaction time results indicated an interference from the other agent's perspective. On those trials where participants had to estimate the likelihood of the actual outcome, it took longer for them to perform their ratings when the other agent could not be certain about the hidden object's identity compared to when she witnessed the same events they did. Although this

finding, in itself, cannot be considered strong evidence for the spontaneous representation of the other agent's hypotheses, as the difference was not significant on trials where participants had to rate the 'possible' outcome, it clearly indicates that they tracked the other agent's knowledge state/certainty, in some manner, and thereby corroborates the results of the rating analyses.

Crucially, despite the fact that it demanded the attribution of a different type of belief content to the agent, Experiment 2 did not differ from Experiment 1 in terms of how much time it took for participants to perform their estimations from their own or from the other agent's perspective. There was no difference in the magnitude of the altercentric bias either, although somewhat more participants demonstrated the predicted effect in Experiment 2 than in Experiment 1.

Such results indicate that it may not be more difficult to attribute beliefs about object identity (and apply disjunctive reasoning to disambiguate it) than to represent other agents' beliefs about object location (and perform disjunctive syllogism in order to infer it). More importantly, it shows that, unlike what the two-system account (Apperly & Butterfill, 2009) and some early studies using dual-identity objects, suggest (see e.g.: Low & Watts, 2013), human adults can and do represent other agents' beliefs about object identity, spontaneously. It is important to note, however, that by the time participants were presented with the response scale, they might have already computed and attributed the belief content to the other agent. In fact, given the findings showing that human adults perform disjunctive syllogism spontaneously, as soon as they receive the information necessary to eliminate the alternatives (see: Cesana-Arlotti et al., 2018), and that participants had to attribute the interim result of their own inferential process to the other (e.g. '*it is either the chick or the frog*'), on trials when the agent did not witness the second box opening, it is rather unlikely that the content of the other agent's mental state was computed at the end, in a retrospective manner, recalling what events the other witnessed and rerunning the whole process. Hence, it might happen that representing the other agent's belief content was actually more difficult in this than in the previous experiment, we just could not capture this difference.

Altogether, findings from Experiment 2 provide further evidence for the representational flexibility of the spontaneous ToM processes and imply that the ability to spontaneously track other agents' inferences may extend to situations in which the observer (and the other agent) has to go through a more complex inferential process to be able draw the appropriate conclusion(s) in the given context. Experiment 3 tested whether indeed this is the case i.e. whether human adults can (and do) track more complex, multi-step inferences of others, in a spontaneous manner.

4.4. Experiment 3

Experiment 3 presented participants with a task which required them to perform multi-step inferences, sometimes from their own, sometimes from the other agent's perspective, by relying on previously acquired rules besides the available perceptual evidence. In particular, participants first had to draw one conclusion, regarding the location of an animal (based on what events they witnessed, using one type of deductive inference) then another conclusion on the basis of the first one, regarding the animal's identity (applying a different logical rule). With such a task, we essentially tested whether adults track others' chains of inferences spontaneously, or only if this is necessary, when they are instructed to do so.

Experiment 3 differed from Experiment 1 in three and from Experiment 2 in two main aspects. First, three different animals could hide in the scene, each of which had its own box, colour-matched to the animal (e.g. the fox always hid only in the red box, the chick in the yellow etc.). Importantly, unlike in Experiment 2, on each trial, only one of the three animals hid, and participants had to infer this hidden animal's identity. Second, participants first learned the conditional rules (such as '*if red box then fox*'), and that this knowledge is shared with the agent, such that they could infer the hidden animal's identity from the information they received about the boxes (which of them are empty/which one remained closed) and could safely assume that the agent can also perform these inferences. Finally, as in Experiment 2, they had to rate the likelihood that a certain animal has hidden at a certain location (in the box that remained closed). Such a design resulted in a situation where the correct solution of the task required participants to combine conditional reasoning with disjunctive inference. Specifically, in order to be able to infer the identity of the hidden animal and attribute the appropriate content to the other, they had to go through the following inferential steps: *(1)'It is neither in Box A nor in Box B, therefore it is in Box C. (2) If it is in Box C then it must be the fox',* in case they engaged in conditional reasoning only after drawing the first conclusion, regarding the location of the animal, or, perform the following inferences, if they applied the appropriate conditional rule at each step of the process: *(1)'It is neither in Box A, therefore it is not the chick' (2) nor in Box B, therefore it is not the frog' (3) 'therefore it must be the fox'*. We hypothesized that if participants track such complex inferences of other agents, they should provide higher ratings for the possible compared to the impossible alternative (i.e. for animals linked with the box that opened second than for the animals linked with the box that opened first) and/or longer RTs on the underspecified belief trials, i.e when the agent represents two equally

likely alternatives regarding where the animal may be and, consequently which animal might have hidden in the scene.

4.4.1 Methods

4.4.2.1 Participants

The final sample included 34 participants M_{age}=28.12, SD_{age}=4.45, 16 males). Participants were recruited via Testable Minds, as before. The selection criteria were identical to the ones used in Experiment 1. All but three participants were right-handed and all of them had at least a high school degree. A further 20 participants were tested but were not included in the analyses: two were excluded because their ratings on the SELF TB actual and/or the SELF TB impossible trials indicated lack of understanding of the task or the use of the rating scale, 11 because they had less than 70% valid trials, four because they failed to answer >50% of the attention check trials correctly (N=4) and three because they failed to meet both inclusion criteria both. The study was approved by the EPKEB United Ethical Committee, Hungary; participants received monetary compensation of 6 USD for their participation. Experiment 2 was preregistered on AsPredicted.org (with the document number 61346); the sample size was determined to match that of Experiment 1.

4.4.2.2 Stimuli and apparatus

The images used in the main task (and in the test trials of the learning phase) were identical to the ones used in Experiment 1 with one exception: the blue box was exchanged for a green one, as it was difficult to find an animal that would have matched in colour to the blue one. Active screen area sizes ranged from 1280 x 720 pixels to 2560 x 1440 pixels (data collected by Testable).

4.4.2.3 Procedure

The general structure of the task was the same as in Experiment 1, except that participants were presented with a learning task after the second training session to teach them the conditional rules. This rule (e.g. '*if yellow box then chicken'*) allowed participants to infer the identity of the hidden animal in the absence of direct visual evidence. This learning phase was approximately 5 minutes long. As a result of this additional phase, the experiment lasted somewhat longer than Experiment 1, for about 35-40 minutes. As in Experiments 1 and 2, each phase was preceded by its own instruction screen (see Supplementary Materials **S4.1.3**), with a maximum time limit set for reading.

Training session 1 and 2

The stimuli and the trial structure of the two training sessions was identical to the ones used in Experiment 1 with two exceptions. First, in the first training session, instead of the kitten, three animals were presented, the ones used in Experiment 2, each of them once, in the box matching their colour. This resulted in three 'animal present' trials (in which participants had to click on the box containing the animal) and three 'animal absent' trials (in which they did not have to do anything). Second, in the question used in the second training session the word *'kitten'* was exchanged for *'animal'* in the question presented above the image (depicting the three closed boxes) at the end of the trial (*'Would the girl search for the animal in?*).

Learning task

The learning task had two phases: a three-trial teaching phase, the purpose of which was to teach participants the conditional rules and highlight that the agent also knows these. In order to evaluate whether participants have learnt these rules, this teaching phase was followed by a six-trial test phase, in which participants had to decide which animal has hidden in the scene depicting one closed and two open boxes, to test whether they have learnt the conditional rules.

Teaching phase trials started with the presentation of a written text (3000 ms) presenting the rules, e.g. '*The frog always hides in the GREEN box. The GIRL also knows this.'*. This was followed by a picture showing the respective animal in the appropriate, colour-matching box, for another 3000 ms, to back up the rule with an example. Participants received altogether three trials, one per animal. After the

third trial, the test phase started. Each test trial started with the presentation of a central black fixation cross, for 500 ms. This was followed by one of the pictures that were used the main task, depicting one closed and two open boxes (with the girl facing the boxes), for a fixed 2500 ms. After this, participants were presented with the picture of the three animals, arranged horizontally, in a fixed order (the frog on the left, the fox in the middle, and the chick on the right). Their task was to click on the animal that they think has hidden in the box that remained closed in the previously presented scene, as fast as possible. The picture of the animals remained on screen for a maximum of 5000 ms or until a response was provided and was followed by detailed feedback, explaining participants which animal should have been selected and why in case their response was not correct. The pictures and the trials were separated by a 250 ISI and ITI, respectively. Just like in the second training session, progress was self-paced to leave enough time for processing the presented text. Altogether participants received six test-trials, with each of the three boxes remaining closed twice, in a fixed pseudorandom order.

Importantly, if participants performed under a pre-set threshold (made more than one error, on the six test phase trials), the whole learning phase was repeated once, to ensure that they start the main task with a firm enough knowledge.

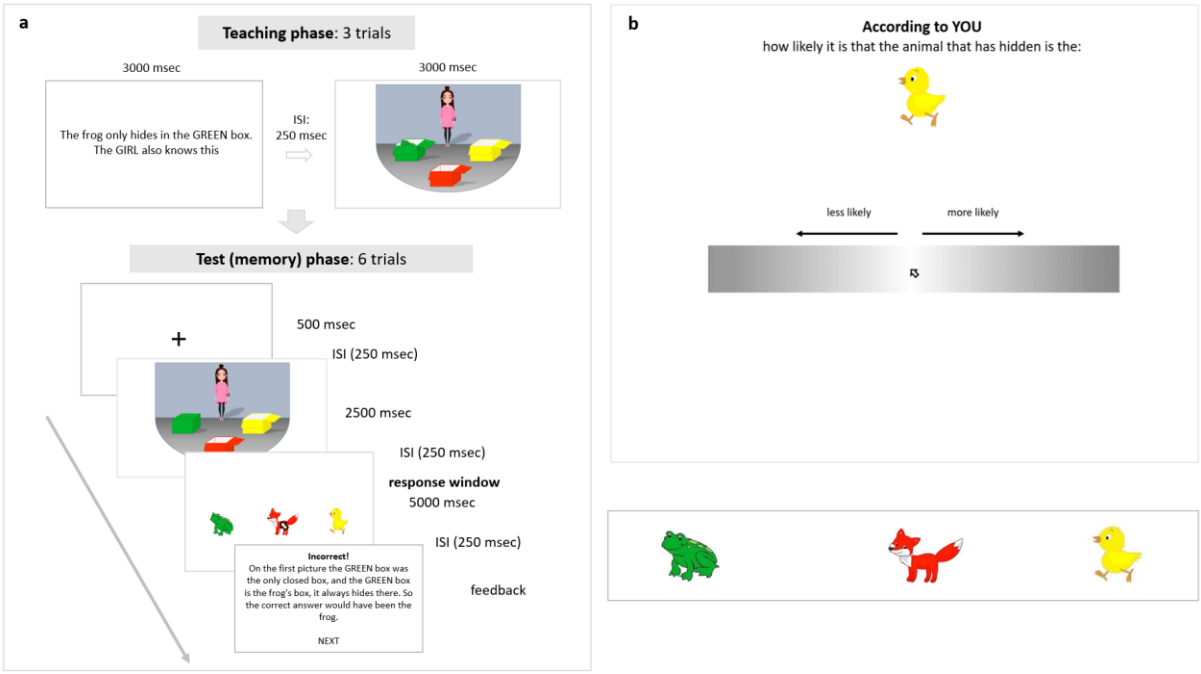


Figure 4.7. (a) The structure of the learning task in Experiment 3. Participants were first taught the conditional rules in three trials and then were tested for their knowledge in six-trials. If they performed under a pre-set threshold, the learning task was repeated once. (b) The response screen presented at the end of the main task, with the to-be-rated animals, below.

Main task

The design and trial structure of the main task was the same as in Experiment 1 with one crucial difference. At the end of the trial, participants had to rate how likely is that a certain animal has been hidden in the scene. To make the use the previously acquired conditional rules necessary, as in Experiment 2, participants were only presented with the picture of one of the three animals below the question (*'According to YOU/the GIRL how likely it is that the animal that has hidden is the'*), without showing them the box with which the animal was paired with. There were two more, minor changes in addition, compared to Experiment 1: First, just like in the first and the second training session, the word *'kitten'* was replaced with *'animal'* in the question of the attention check trials. Second, as in Experiment 2, participants were required less precision when clicking at the fixation cross than in Experiment 1. Everything else, including the experimental conditions, the counterbalancing and the rules of pseudorandomization were exactly the same as in Experiment 1, with the alternative type defined by which box the to-be-rated animal was linked with (the one that opened first – impossible; the one that opened second – possible or the one that remained closed at the end – actual).

4.4.2.4 Data analysis

The data was analysed as before, using the same data and participant exclusion criteria and running the same analyses as in Experiment 2 (see: Section 4.3.1.4: Data analysis).

4.4.3 Results

4.4.3.1 Rating analyses

Other-perspective trials

To investigate whether participants are able to track more complex inferences of another agent, when they are required to monitor what conclusions the other might draw in the given situation, first we analysed the likelihood estimations they provided on the other-perspective trials.

The pattern of results was similar to the one observed in Experiment 1: Friedman tests indicated a significant overall difference between the three types of alternatives, on both the true and underspecified belief trials (TB: $\chi^2(2)=51.2$, $p< .001$, Kendall's $W=0.753$; UB: $\chi^2(2)=49.94$, $p< .001$, Kendall's $W=0.734$). As in Experiment 1, on the true belief trials, participants' ratings were significantly higher for the actual than for either of the other two alternatives (actual-possible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; actual-impossible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$), while ratings for possible and impossible alternative did not differ significantly from each other (possible-impossible: $Z=-1.29$, $p_{adj(3)}= .197$, $r=0.221$; see **Figure 4.8**). This suggests that participants understood that, when the other agent saw both box-openings, she could infer the identity of the hidden animal, just like they could, by applying the same logical rules. On trials where the agent did not witness the second opening (thus could not be certain about the identity of the hidden animal), participants' ratings were similar for the actual and the possible alternative: they were significantly higher than ratings for the impossible alternative (actual-impossible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; possible-impossible: $Z=-5.00$, $p_{adj(3)}< .001$, $r=0.858$) and cumulated around 0.5-0.6, just like in Experiment 1. Furthermore, a clear egocentric bias also emerged in the ratings: participants considered it significantly more likely that, according to the agent, the hidden animal was the one that was actually hidden in the closed box than that it was the one only the other agent considered a possible alternative ($M_{actual}=0.62$, $SD_{actual}=0.13$; $M_{possible}=0.53$, $SD_{possible}=0.18$; $Z=-3.17$, $p_{adj(3)}= .002$, $r=0.544$), despite the two alternatives were equally likely for the agent, suggesting difficulties with inhibiting their own perspective.

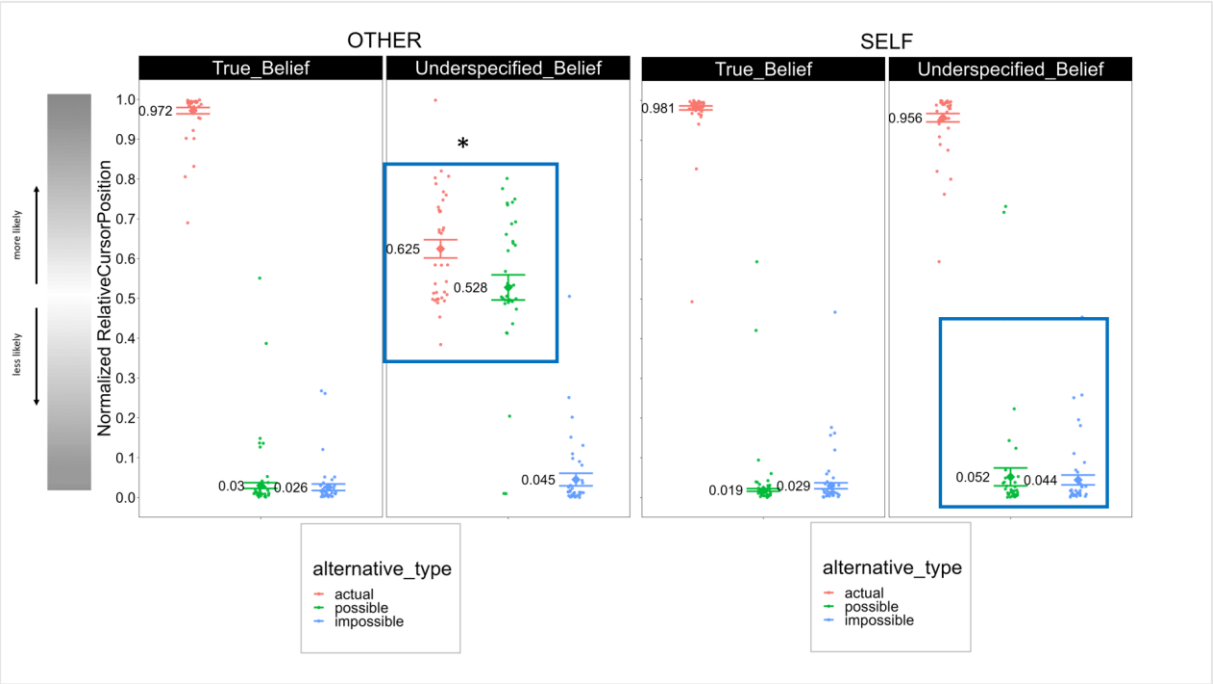


Figure 4.8 Mean ratings (mean normalized relative cursor position) on the self- and other perspective trials per belief and alternative type in Experiment 3. Higher values indicate that participants considered it more likely that the presented animal was the one hidden in the scene, either from their own (SELF trials) or from the agent's perspective (OTHER trials). Error bars represent 95% CI, dots show the individual means. Blue frames indicate our two main foci of interest: the difference between the possible and impossible alternative on the self-perspective and the actual and possible alternative on the other-perspective underspecified belief trials The first comparison investigates whether there is an altercentric bias in the estimations (indicating spontaneous representation of what the other considers possible) on the self-perspective trials, the second test for the explicit representation of the two alternatives the agent could represent on the other-perspective trials. Stars indicate significant differences between the two experimental conditions. *: p<0.05, **: p<0.01

Self-perspective trials

To test whether participants themselves performed the necessary computations (inferred the identity of the hidden animal) on the one hand, and our main hypothesis, that they tracked the inferences of the other agent, also when this was not necessary, on the other, next we analysed ratings on the self-perspective trials. Analyses revealed a significant overall difference between the three alternatives on both the true and the underspecified trials (TB: $\chi^2(2)=51.06$, $p< .001$, Kendall's $W=0.751$; UB: $\chi^2(2)=52.58$, $p< .001$, Kendall's $W=0.773$). Mirroring the findings of Experiment 1, ratings were significantly higher for the actual than for either the possible or the impossible alternative, on both the true (actual-possible: $Z=-5.09$, $p_{adj(3)}< .001$, $r=0.873$; actual-impossible: $Z=-5.09$, $p_{adj(3)}< 001.$, $r=0.873$)

and the underspecified belief trials (actual-possible: Z=-5.09, $p_{adj(3)}$< .001, r=0.873; actual-impossible: Z=-5.09, $p_{adj(3)}$< .001, r=0.873), indicating that participants were able to track the events and infer the identity of the hidden animal.

Crucially, however, there was no sign of the altercentric bias observed in the previous experiments: participants' ratings were similar for the possible (i.e. for the animal linked with the box that opened second) and for the impossible alternative (i.e. for the animal linked with the box that opened first), not only on those trials where the other agent had the same knowledge they did (Z=-1.27, $p_{adj(3)}$= .203, r=0.218) but also on those where the agent did not witness the second box opening, therefore ended up representing two, equally likely alternatives, regarding the hidden animal's location and identity ($M_{possible}$=0.05, $SD_{possible}$=0.13 versus $M_{impossible}$=0.04, $SD_{impossible}$=0.07; Z=-1.14, $p_{adj(3)}$= .257, r=0.196). Inspection of the individual means revealed that the predicted effect was present in only 7 out of the 36 participants (using the stricter criteria), and there were altogether N=13 participant who demonstrated a rating pattern that was in line with our hypotheses (i.e. SELF UB difference score minus SELF TB difference score >0), according to the more lenient criteria. Excluding those participants who were not included in the RT analyses below (N=2), did not change the main results (SELF true belief trials – actual-impossible: Z=-4.94, $p_{adj(3)}$< .001, r=0.873; SELF underspecified belief trials - possible-impossible: Z=-1.29, $p_{adj(3)}$= .196, r=0.228; OTHER underspecified belief trials – impossible-possible: Z=-4.87, $p_{adj(3)}$< .001, r=0.860; actual-possible: Z=-3.07, $p_{adj(3)}$= .002, r=0.542).[24]

4.4.3.2 Reaction time analyses

Other-perspective trials

To investigate whether the egocentric bias, present in participants' likelihood estimations, also emerges in their reaction times, on those trials where they have to perform the rating from the other agent's perspective and to provide further evidence for the representation of both alternatives the other agent did, in these cases, next we analysed participants' RTs of the other-perspective trials. A 2 x 3 repeated measures ANOVA, with belief (TB versus UB) and alternative type (actual, possible, impossible) as within-subject factors, yielded a significant main effect of belief (F(1, 31)=15.87, p< .001, η_p^2= .339), resulting from the generally higher reaction times on the underspecified than on the true belief trials, a significant main effect of alternative type (F(2, 62)=4.53, p= .015, η_p^2= .128), as well as a

[24] The only person who could be considered an outlier based on his mean rating on the self-perspective UB possible trials, was already excluded based on his mean rating on the self-perspective TB actual trials.

significant belief x alternative type interaction ($F(2, 62)=6.61$, $p= .003$, $\eta_p^2= .176$), reflecting the marked difference in the reaction time pattern of the two types of belief trials (see **Figure 4.9**). Specifically, on the true belief trials, i.e. when the other agent could infer the hidden animal's identity, participants were faster to rate the actual than either the possible or the impossible alternative, from the other agent's perspective, with RTs being the longest for the possible alternative. On the underspecified belief trials, where the other agent represented two, equally likely alternatives, they were equally fast in performing their ratings on the actual and the possible trials. Pairwise comparisons, run separately for the true and underspecified belief trials, revealed a marginally significant difference between the actual and the possible alternative on the true belief trials ($t(31)=-2.35.$, $p_{adj(3)}= .078$, $d=0.415$), and significant difference between impossible and the other two alternatives (actual-impossible: $t(31)=-3.04$, $p_{adj(3)}= .015$, $d=0.566$; possible-impossible: ($t(31)=-2.84$, $p_{adj(3)}= .016$, $d=0.502$) on the underspecified belief trials, with no difference between the time necessary to rate the actual and the possible alternative ($t(31)=-0.10.$, $p_{adj(3)}= .924$, $d=0.017$). This latter finding, i.e. that it took roughly equal time for participants to perform their judgements for the actual and the possible alternative, from the agent's perspective, provides further evidence that they understood that when the agent did not witness the second box-opening she represented two equally likely alternatives.

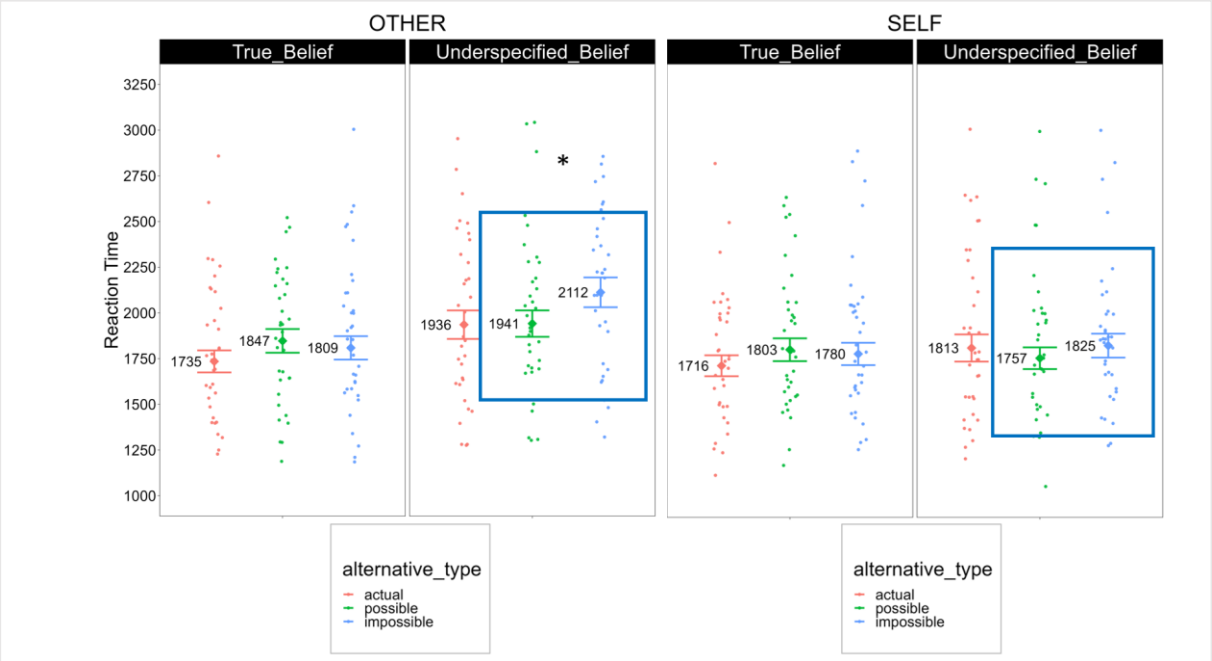


Figure 4.9. Mean reaction times (time necessary to perform the likelihood estimations) on the self- and other perspective trials per belief and alternative type in Experiment 3. Error bars represent 95% CI, dots show the individual means. Blue fames indicate our two main foci of interest: the difference between mean rating time of the possible and impossible alternative on the self- and on the other-perspective underspecified belief trials. *: p<0.05, **: p<0.01

Self-perspective trials

To investigate whether we find evidence in support of our main hypothesis in the RTs (i.e. to investigate whether there is *any* sign that participants spontaneously tracked the other agent's inferences), as in the previous two experiments, we also analysed the time it took participants to perform their ratings on the self-perspective trials, running the same analyses, we performed above, on the other-perspective trials.

A 2 x 3 repeated-measures ANOVA, with belief (TB versus UB) and alternative type (actual, possible, impossible) as within-subject factors, yielded no significant main effect of belief ($F(1, 31)=0.93$, $p= .33$, $\eta_p^2= .029$) or alternative type ($F(2, 62)=2.13$, $p= .128$, $\eta_p^2= .064$). There was, however, a significant belief x alternative type interaction ($F(1, 62)=3.39$, $p= .040$, $\eta_p^2= .099$). As can be seen on the figure, when the agent had the same knowledge as participants did, thus could arrive at the same conclusion, the pattern was similar to the one observed on the other-perspective trials: participants' responses were faster for the actual alternative than either for the possible or the impossible alternative, with no difference between the latter two. Pairwise comparisons indicated a significant difference between the actual and the possible ($t(31)=-2.90$, $p_{adj(3)}= .021$, $d=0.513$) and a marginally significant difference between the actual and the impossible alternative, after adjusting for multiple comparisons ($t(31)=2.29$, $p_{adj(3)}= .058$, $d=0.405$).

Regarding the underspecified belief trials, in line with the results obtained in the rating analyses, there was no sign of altercentric interference from the other agent's perspective: participants were actually somewhat faster in rating the possible alternative, i.e. the animal linked with the box that opened second, than providing their ratings either for the actual or for the impossible alternative, although the difference was not significant (actual-possible: $t(31)=0.68$, $p_{adj(3)}= .503$, $d=0.120$; possible-impossible: $t(31)=-1.60$, $p_{adj(3)}= .363$, d=0.282).

4.4.3.3 Comparison of Experiment 1 and Experiment 3

To further investigate how the increase in the complexity of the inferences participants had to make (both from first- and third-person perspective), in this compared to the first experiment, affected the tracking of the other agent's inferences, and to better understand what may explain the lack of the predicted effect in the current experiment, we compared the two experiments: 1) along the difference scores computed from the mean ratings provided for (a) the possible and impossible alternative of the SELF underspecified belief and (b) the possible and impossible as well as the actual and possible

alternative of the OTHER underspecified belief trials, and 2) along the mean reaction times necessary to perform these ratings.

Rating analyses

Reflecting the differences in the findings, pairwise comparisons revealed a significant difference between the two experiments in terms of the magnitude of the altercentric bias, observed on the SELF underspecified belief trials: it was larger in Experiment 1 than in Experiment 3 (SELF UB possible-impossible difference score: U=386.00, $p_{adj(3)}$= .036, r=0.302).

Although the egocentric bias was more pronounced in this third than in the first experiment, after adjusting for multiple comparisons, the difference between the two experiments was not significant (OTHER UB actual-possible difference score: U=442.00, p_{unadj}=.066, $p_{adj(3)}$= .132, r=0.221). There was no significant difference between the two experiments in the magnitude of the OTHER UB possible-impossible difference score either (U=534.00, p_{unadj}= .464, r=0.088).

Despite the different pattern – the presence of the predicted effect in Experiment 1 and the absence of it in Experiment 3 - the number of participants demonstrating the effect was roughly the same in both (N_{Exp3}=7 versus N_{Exp1}=8, $\chi^2(2, N=69)=0.52$, p= .819, Cramer's V=0.028), according to the stricter criteria. There were, however, altogether more participants in Experiment 1 than in Experiment 3 whose corrected difference score on the SELF UB trials was in line with our predictions, according to the more leaner criteria (N_{Exp3}=13 versus N_{Exp1}=20, $\chi^2(2, N=69)=3.31$, p= .069, Cramer's V=0.219).

Reaction time analyses

To capture potential differences in the processes underlying participants' likelihood estimations in the two experiments, as a final step, we ran two 2 x 2 x 3 mixed ANOVAs, with experiment as a between-subject and belief and alternative type as within-subject factors, separately for the self- and other-perspective trials, on the log-transformed RT data.

Analysis of the other perspective trials yielded a significant main effect of belief ($F(1, 63)=31.98$, $p<$ .001, η_p^2= .337), alternative type ($F(1, 126)=5.95$, p= .003, η_p^2= .086) and a significant belief x alternative type interaction ($F(1, 126)=3.18$, p= .045, η_p^2= .048), as well as a significant main effect of experiment ($F(1, 63)=8.59$, p= .005, η_p^2= .120). As can be seen on the figure, contrary to what one would expect, reaction times were *lower* in Experiment 3 than in Experiment 1, independent of the alternative

participants had to rate and the belief of the agent. Pairwise comparisons confirmed that in Experiment 3 it took significantly less time for participants to estimate the likelihood of the presented alternative from the agent's perspective than in Experiment 1, in all but one experimental condition (TB - actual: $t(63)=3.06$, $p_{adj(3)=}$ .003, d=0.645; possible: Welch's $t(57.37)=2.16$, $p_{adj(3)=}$ .035, d=0.532; impossible: Welch's $t(56.78)=2.75$, $p_{adj(3)=}$ .008, d=0.668; UB – actual: $t(49.22)=2.58$, $p_{adj(3)=}$ .013, d=0.737; possible: $t(63)=2.65$, $p_{adj(3)=}$ .010, d=0.742; impossible: Welch's $t(53.24)=1.19$, $p_{adj(3)=}$ .240, d=0.256). Despite the markedly different reaction time patterns in the two experiments, neither the experiment x belief ($F(1, 63)=0.35$, p= .559, η_p^2= .005) nor the experiment x alternative type interaction was significant ($F(2, 16)=1.21$, p= .302, η_p^2= .019). There was no significant experiment x belief x alternative type interaction either ($F(2, 126)=1.90$, p= .153, η_p^2= .029).

Analysis of the self perspective trials revealed a tendency level main effect of alternative type ($F(1.75, 110.18)=2.47$, p= .097, η_p^2= .038) but no significant main effect of belief ($F(1, 63)=1.31$, p= .257, η_p^2= .020), significant belief x alternative type interaction ($F(2, 126)=0.89$, p= .415, η_p^2= .014). There was no significant experiment x alternative type ($F(1.75, 110.18)=1.97$, p= .149, η_p^2= .030) and experiment x belief interaction either ($F(1, 63)=0.09$, p= .767, η_p^2= .001). There was, however, a significant main effect of experiment ($F(1, 63)=11.9$, p= .001, η_p^2= .160), resulting from the fact that, despite the larger complexity of the inferential process participants had to go through in order to provide the appropriate response in this than in the first experiment, it took less time for them to perform their ratings, for all three alternatives, independent of the agent's belief (TB - actual: $t(59.30)=3.74$, $p_{adj(3)}$< .001, d=0.966; possible: $t(63)=2.79$, $p_{adj(3)}$= .010, d=0.663; impossible: $t(63)=2.89$, $p_{adj(3)}$= .010, d=0.681; UB – actual: $t(50.56)=2.71$, $p_{adj(3)}$= .018, d=0.660; possible: $t(63)=4.10$, $p_{adj(3)}$< .001, d=0*962*; impossible: $t(63)=2.32$, $p_{adj(3)}$ .024, d=0.576). In addition, there was also a marginally significant experiment x belief x alternative type interaction ($F(1.75, 110.18)=2.99$, p= .055, η_p^2= .045), reflecting the different reaction time pattern in the two experiments on the SELF true belief trials.

4.4.3 Discussion

The results of Experiment 3 show that, when they have to explicitly track what the agent may think or know, adults can monitor what conclusions another agent may draw from the beliefs she holds, across multiple, consecutive inferences (where the outcome of one inference serves as the premise of the subsequent one). On those trials where participants had to take the agent's perspective and the agent did not witness the second box opening, ratings were similar for the animals linked with the box that

opened second and for the animals linked with the box that remained closed at the end, indicating that participants represented both alternatives the agent could represent regarding the identity of the hidden animal. This does not mean that the ratings on the OTHER trials were the same in Experiment 3 as in the previous experiments. With the increase in the complexity of the inference participants had to perform, the egocentric bias observed in Experiment 2 (and to some extent already in Experiment 1) became much more pronounced in Experiment 3, corroborating previous studies which demonstrated a similar impact of cognitive load on adults' explicit perspective-taking abilities (see e.g. Qureshi et al., 2010).

Importantly, however, regarding our main hypothesis, there was no sign of the altercentric estimation bias that was present on the SELF underspecified belief trials in the previous two experiments. Ratings for the possible and the impossible alternative did not differ significantly. There was no sign of altercentric interference from the other agent's perspective in the reaction times either. Such a finding points to the possible limitations of adults' capacity to represent the conclusions other agents may draw in a situation, by showing that spontaneous tracking is hindered by the growing complexity of the inferences participants have to perform.

Notably, besides the lack of altercentric effect, reaction times were also much shorter in Experiment 3 than in Experiment 1 (or Experiment 2), in general, suggesting that participants may not have tracked the other agent's beliefs at all on the SELF trials and may have relied on completely different cognitive processes to solve the task in this experiment than in the first one (both on the SELF and on the OTHER trials). Specifically, the increased task difficulty may have led participants to suspend the tracking of the other agent's beliefs, when this was not strictly necessary to provide an answer, as all the available executive resources had to be focused on the actual task. Furthermore, it may have led to the use of shortcut strategies, which might explain why RTs were so much shorter in Experiment 3 than they were in Experiment 1. For instance, participants might have only paid attention to the colour of the boxes, and then provided responses simply on the basis of the colours. On the OTHER underspecified belief trials, they might have memorized the colour of the closed boxes while the agent was looking at those (e.g, '*red and green'*) and then matched the labels to the colour of the animal presented on the response screen, instead of engaging in proper mentalizing and attributing the propositions *'The animal is either in the green or the red box. Therefore, it is either the frog or the fox'* to the other.

Alternatively, shorter reaction times might have resulted from performing the inferences (for the other agent) at an earlier timepoint in this than in the first or the second experiment, both when participants had to take the other's perspective and when this was not necessary. It may happen that, while in the previous experiments, participants drew the conclusion from the other agent's perspective only at the end of the trial, when they received the test question in this experiment they did this when the agent turned away, to counteract the difficulty of the task they had to perform.

It is important to note that Experiment 3 differed from Experiment 1 not just in the complexity of the inferences participants had to perform from first-person (and, on the explicit trials, from third-person) perspective, more specifically, in the number of inferential steps they had to make and logical rules they had to apply, but also in two other aspects. First, tracking of the other agent's inferences required participants to attribute her two (and not just one) sets of alternatives on the underspecified belief trials: first one about the potential location ('*it is either in the green or the red box*') and then another one about the potential identity of the hidden animal (therefore, '*it is either the frog or the fox'*). Participants either had to perform the extra inferential step and combine the output of the disjunctive syllogism with the appropriate conditional rule each time a box opened or at the end of the trial, in a retrospective manner, recalling what the other had and had not witnessed. It may happen that participants actually tracked the other agent's inferences spontaneously, but only about the object's (potential) locations, and simply did not perform the final inferential step, regarding the potential identity of the animal. Second, the task did not only require participants to attribute two sets of alternatives on the underspecified belief trials but also to compute the content of the to-be-attributed beliefs on a trial-by-trial basis, continuously tracking which box the other agent saw empty to be able to identify what alternatives the agent considers on these trials.

To investigate whether human adults track multi-step inferences of other agents, spontaneously, if this task is less demanding, we ran a fourth experiment, with a simplified version of the task. Experiment 4 also required participants to track multiple, consecutive inferences of another agent, however, they did not have to compute a new content on each and every underspecified belief trial, as the alternatives the other agent represented in these cases, were constant throughout the task. Thus, Experiment 4 essentially tested whether it is easier to spontaneously track multi-step inferences of other agents when the final conclusion they may draw is based on their stable prior beliefs, than to track multi-step inferences when those are based on computations performed on the spot (based on momentarily available information), in those cases where the agents end up representing multiple alternatives.

4.5. Experiment 4

Experiment 4 investigated whether human adults spontaneously track multiple, consecutive inferences of other agents, that are based on the agents' stable prior beliefs in situations where the task-relevant inference requires participants to combine conditional reasoning with disjunctive

inference and to draw first a conclusion about one characteristic of an object (location) before being able to make inference about another (identity).

The task had a similar logic as the one used in Experiment 3 did, however, it did not require participants to continuously track what events the other agent had and had not witnessed to be able to attribute the appropriate belief content, i.e. to correctly represent what conclusions the other agent may draw regarding the identity of the hidden animal. Instead, they had to take into account a stable belief[25] of the agent regarding what is possible in a certain context (that differed from what they themselves considered possible), to identify what alternatives she considers (about the identity of the animal) on those trials where she could not unambiguously infer which animal had hidden in the closed box. We assumed that the fact that participants have to take into account a stable belief of the other agent, to represent the conclusions she may draw regarding the identity of the hidden animal, would make the task easier for them. If so, then this would increase the likelihood that, in case human adults are indeed able to track multi-step inferences of other agents, participants would spontaneously represent both conclusions the other may draw in the given context, not just the one that can be drawn from the perceptual evidence (about the location of the hidden animal) but also the one that follows from the first conclusion (what this animal is).

To test whether this is the case we designed a task that mimicked a hide-and-seek scenario where there are, for instance, three hiding locations, two hiders and a seeker, and at each hiding round only one hider can hide, but it is not known in advance which one. The seeker can unequivocally infer the location of the hider (e.g. the top of the tree), after excluding the two other options, however, he cannot be certain about the hider's identity, as he mistakenly believes that both hiders can climb up high, while, in reality, only one of them is able to. If an observer tracks the multi-step inferences of the seeker, she must represent the two alternatives the seeker considers (that either of the two hiders can be on the top of the tree), even if she knows who can, in fact, climb so high, and thus, can disambiguate the identity of the hider.

Analogously, in the current study, we presented participants with a task, involving three hiding locations (boxes) and three animals, in which an agent witnessed all the events they did, hence, just like participants, could unambiguously infer the *location, but not necessarily the identity* of the hidden animal, after two of the boxes were revealed to be empty on every trial. To create a situation in which the agent holds an underspecified belief in some cases, we manipulated what information the agent and the participant received a priori, such that the agent and the participant received different

[25] Here we use the term stable belief to refer to a task-relevant belief of the agent the content of which was computed before the task and remained constant throughout the trials which differ from long-lasting beliefs (such as political or religious beliefs), that influence others' inferences not only in the context of the specific task but also outside the lab.

information on whether one of the animals can hide in two boxes or only one box. In particular, participants were first taught three conditional rules, according to which animals hide based on their colour (such as 'the red animal hides in the red box'), as in Experiment 3. However, this time they were told that one of the animals is bicolored (a green turtle, that had yellow spots) and thus it can hide in two boxes on the basis of its colour (in the green and the yellow box), emphasizing that the agent is also aware of these rules. Then they were informed that this specific animal can actually hide only in the green box (as the yellow spots were just painted) but, importantly, this is something the agent is not aware of. As a result, the agent had an incorrect belief about where the turtle can hide ('*the turtle can hide in both the green and the yellow box*'), which led her to mistakenly believe that two of the animals (the turtle or the chick) can hide in the yellow box, and, consequently, to represent two alternatives regarding the identity of the hidden animal, on those trials where this specific box remained closed at the end. Participants had to take into account the incorrect belief of the other agent, more specifically, the fact that she applies a different conditional rule than they do, on these trials (*'if it is the yellow box then it is either the chick or the turtle'* instead of '*if it is the yellow box than it is the chick'*), to be able to correctly represent what conclusions the other may draw regarding the hidden animal's identity (from the beliefs she holds about its location).

Notably, our manipulation meant that in this experiment, 'belief type' was determined by which box remained closed at the end of the trial (and not what events the other agent witnessed): (i) the yellow box, about which the agent mistakenly believed it can contain either the chick or the turtle, (underspecified belief trials); (ii) the green box, which was linked with the turtle, an animal the agent also linked with another box (true belief[26]-ambiguous trials); or (iii) the red box, which was linked with only one animal, both by participants and the agent (true belief-unambiguous trials). We hypothesized that if participants spontaneously represent what conclusions the other agent might draw in the above-described situation (i.e. can and do track multiple, consecutive inferences of other agents, if the appropriate attributions are less demanding than they were in the previous experiment given that they rely on a stable, a priori established belief of the agent), the altercentric bias should re-emerge in this experiment, on the self-perspective trials. In specific, participants' ratings should be higher and/or take longer on those SELF underspecified belief trials, where the to-be-rated animal is the one about which the agent mistakenly believes it can hide in the yellow box (it is the turtle - 'possible' alternative) compared to those trials where it is the animal about which no one believes that it can hide there (it is the fox - 'impossible' alternative). In addition, we conjectured that, as the to-be-attributed

[26] In the preregistration we used the uncertain-certain terminology, instead of the underspecified-true belief terms. We changed it in this chapter to be consistent throughout the thesis and also for reasons already mentioned in Chapter 3 (i.e. 'uncertain' would be a rather misleading label as participants themselves are not uncertain regarding the content of the agent's belief).

underspecified belief content is constant throughout the task, it might be encoded in more detail, such that it may even include information about the probability the agent assigns to the two alternatives (that there is a 50% probability that the chick and a 50% probability that the turtle is hiding in the yellow box, according to her). We hypothesized that if participants spontaneously represent the probabilities from the other agent's perspective (that the yellow box contains either the chick or the turtle, each with 50% probability) then, on the SELF underspecified belief trials, this might be reflected not only in how they rate the likelihood of the 'possible' alternative (that the turtle is in the yellow box) but possibly even in how they rate the actually hidden animal (that the chick is in the yellow box), i.e. it would bias estimations not only towards higher ratings for the turtle but also towards lower ratings for the chick. In particular, we hypothesized that, if participants also encode the probabilities the other agent assigns to the two alternatives, their ratings should be lower for the actually hidden animal on the SELF underspecified belief than on the SELF true belief unambiguous trials (i.e. for the chick on those trials where the yellow box remains closed than for the fox on those trials where the red box remains closed at the end).

4.5.1 Methods

4.5.1.1 Participants

The final sample consisted of 35 participants M_{age}=25.97, SD_{age}=5.46, 16 males, 2 'other'). Participants were recruited via Testable Minds as in the previous experiments. The selection criteria were identical to the ones used in Experiment 1-3. All of the participants had at least a high school degree; 4 were left-handed, 1 was ambidextrous, the rest ere right-handed. Further 16 participants were tested but their data was not analysed: one participant was excluded because his ratings on the SELF true belief unambiguous trials indicated a lack of understanding of the task or the use of the rating scale (for details of the exclusion criteria used in this study, see Section 4.5.3: Data analysis), 7 because they failed to answer >50% of the attention check trials correctly, 8 because they failed to meet both participant inclusion criteria (besides having <50% correct answer on the check questions they also had less than 70% valid trials). The study was approved by the EPKEB United Ethical Committee, Hungary; participants received monetary compensation of 6.8 USD for their participation. Experiment

4 was preregistered on AsPredicted.org (with the document number 6585); the sample size was determined to match that of the previous experiment.

4.5.1.2 Stimuli and apparatus

The images used in the experimental trials of the main task were the same as in Experiment 3, with one crucial exception: the girl was always facing the boxes, hence she always had the same knowledge as the participant. The stimuli used on the attention check trials were identical to the ones used in Experiment 2, except that the image of the frog was replaced by that of a turtle, with yellow patches on its back. The same image was used to collect responses in the test trials of the learning phase. Active screen area sizes ranged from 1280x720 pixels to 2560 x 1440 pixels (data automatically collected by Testable).

4.5.1.3 Procedure

The general structure of the task differed from that of Experiment 3, in three ways: (1) participants were presented with the first training session only (see below); (2) the learning task included a one-trial belief correction phase (that resulted in the participant and the agent holding different beliefs regarding where one animal could hide in the scene) and (3) there were extra questions at the end of each test trial of the learning phase to check whether participants understood when the agent can and cannot be certain about the hidden animals identity. In addition, to be able to present the same number of trials in all experimental conditions, as before, despite having three types of beliefs instead of just two, the main task included 108 (instead of 72) test trials. As a result, Experiment 4 lasted longer than Experiment 3, for about 40-45 minutes. Just like before, each phase was preceded by its own instruction screen (see Supplementary Materials **S4.1.4**), with a maximum time limit set, to ensure that participants do not disengage from the task for long periods and finish it within a maximum of 45 minutes.

Training session

The training session was identical to the first training session of Experiment 3, with the frog being replaced by the turtle on the trial where the green box remained open, and the animal was present in the box.

Learning task

The learning task was made up of three phases, with the first and the third having its own instruction screen (see the Supplementary Materials): a three-trial teaching phase, teaching participants (as well as the agent) that, unlike the other two animals, the turtle can hide in two boxes; a one-trial belief correction phase, informing participants that the turtle can actually hide only in one box and this is something the agent does not know; and a nine-trial test phase, with trial-by-trial feedback, in which participants had to decide which of the three animals has hidden in the previously presented scene, depicting one closed and two open boxes, the first from self- then from other-perspective, to test whether (i) they have learnt the conditional rules and understood (ii) that the agent incorrectly believes that two of the animals can hide in one of the boxes (either the chick or the turtle), as a result of entertaining an incorrect belief regarding where the turtle can hide.

The stimuli and the trial structure of the teaching phase is presented on **Figure 4.10**. Teaching phase trials started with the presentation of a written text for 3000 ms, as before, that presented the conditional rule, changing the previously used '*always hides'* phrase to '*can only hide'* for the fox and the chick to match the expression used for the turtle ('*The turtle can hide both in the GREEN and the YELLOW box.'*) as closely as possible. The written text was followed by a picture showing the respective animal in the appropriate, colour-matching box, for another 3000 ms, to back up the rule, with the turtle presented first in the green and then in the yellow box (altogether for 6000 ms). Participants received three teaching trials, one per animal. Then the belief correction phase followed: participants were presented first with a picture of the turtle, showing its all-green belly, for 7000 ms, along with a text informing them that the yellow patches are only painted on the turtle's back, but the animal is in fact completely green. Then they were presented with a text telling them that, consequently, the turtle can only hide in the green box, again for 7000 ms. Crucially, in both cases, it was highly emphasized that this is something the agent does not know. After the second screen, the turtle was presented once

again, in the green box, in the absence of the agent, for 3000 ms. Following this extra trial, participants progressed to the test phase of the learning task.

The structure of the learning-task test phase trials was the same as before, except that participants had to provide two responses at the end. First, they had to indicate what they think, which of the three presented animals has hidden in the previous scene. Then, after selecting one animal and receiving feedback, the same way as in Experiment 3, they were presented with the picture of the three animals again, asking them '*What would the girl respond to the question*'. They had to answer by clicking on one or two animals, depending on the belief of the agent, with the picture of the animals remaining on the screen until they provided two responses on those trials where the agent represented two alternatives (or for a maximum of 5000 ms). The pictures and the trials were separated by a 250 ISI and ITI, respectively. Altogether participants received nine test trials, with the yellow box remaining closed four, the green three, and the red box two times, in a fixed pseudorandom order. As in Experiment 3, if participants performed under a pre-set threshold (70%), either on the self- or on the other-perspective trials and/or under 50% on those trials where the closed box was the one about which the agent believed that it could serve as a hiding place for two animals, the learning phase was repeated once.

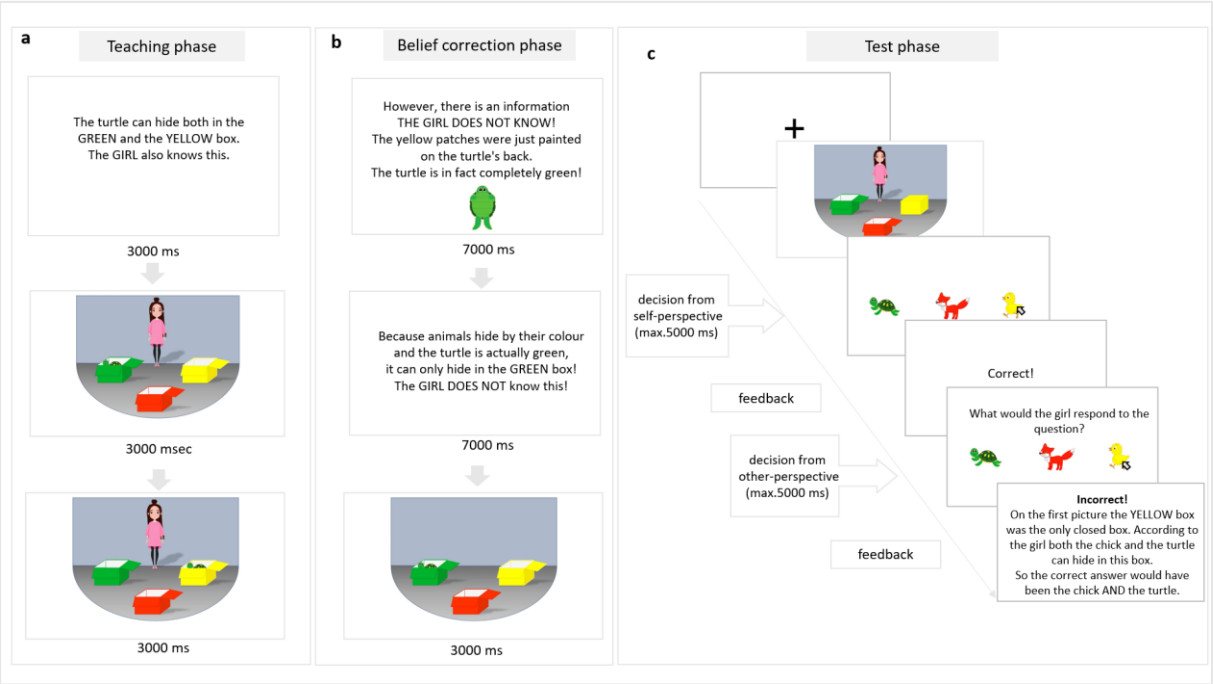


Figure 4.10. (a-c) The structure of the learning task in Experiment 4. Participants were first taught conditional rules in three trials, learning that one of the animals (the turtle) can hide in two boxes. Then, in an additional trial, they were informed that this animal actually hides in only one box (the green) and the agent does not know this. Then they received nine test trials, where they had to decide which animal has hidden in the presented scene first from their own and then from the agent's perspective to test (i) their knowledge and (ii) whether they understood that when the yellow box remains closed the agent represents two alternatives.

Main Task

The trial structure of the experimental trials was similar to the one used in Experiment 3, except that the picture-sequence, following the perspective prompt, was made up of only three elements: one picture with all three boxes closed, one with one and one with two boxes open, with the agent always facing the boxes, presented in this order (see **Figure 4.11** for details).

Based on the colour of the box that remained closed experimental trials could be either:

1) underspecified belief trials, when the yellow box remained closed, i.e. the box linked with two animals by the agent
2) true belief – unambiguous trials, when the red box remained closed, i.e. the box which was linked with an animal not linked with any other box, or
3) true belief-ambiguous trials, when the closed box was the green one, i.e. a box linked with an animal that was linked with two boxes by the agent.

We will use the UB, TB_UNAMB and TB_AMB abbreviations to refer to these three belief types in the followings, respectively. Underspecified belief trials could be either: possible, actual or impossible trials, depending on whether the to-be-rated animal was the one linked with the yellow box only by the agent, but not by the participant (turtle), both by the participant and the agent (chick) or by none of them (fox), respectively. For true belief trials, the alternative type was defined by whether the to-be-rated animal was the one linked with the box that remained closed at the end (actual alternative, e.g. the fox on the true belief unambiguous - 'red box' trials) or not. Therefore, on these trials, besides the actual there were two impossible alternatives (i.e. the chick and the turtle on the true belief unambiguous – 'red box', – and the chick and the fox on the true belief ambiguous – 'green box' - trials; see **Figure 4.11** for an illustration).[27] As such the experiment followed a 2 (SELF or OTHER perspective) x3 (belief type: true-unambiguous, true-ambiguous, underspecified belief) x 3 (alternative type: actual, possible, impossible) design.

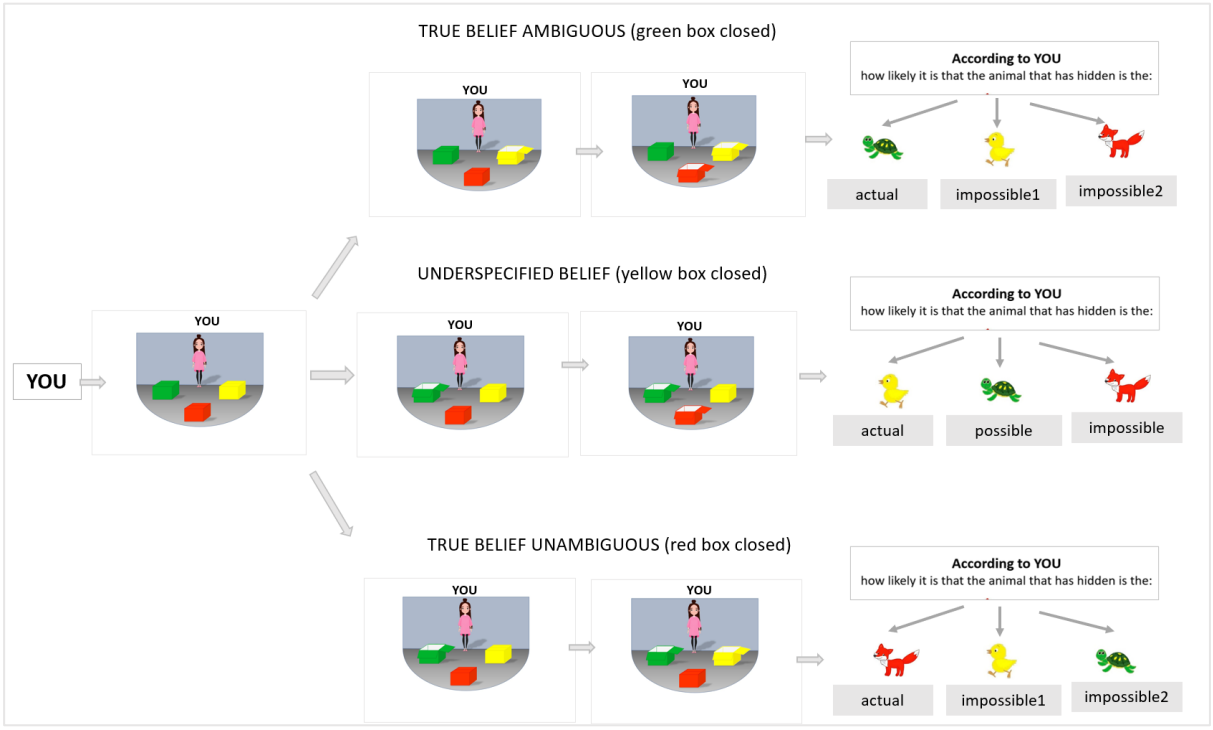


Figure 4.11. Schematic illustration of the SELF true ambiguous (upper panel) underspecified belief (middle panel) and true unambiguous belief trials (lower panel) in the main task of Experiment 4, with the three types of alternatives participants had to rate (actual, possible and impossible on the underspecified and actual, impossible1 and impossible2 on the true belief trials). The trial structure was similar to the one used in Experiment 3, but the agent witnessed all the events, hence fewer events were presented, and belief type was determined which box remained closed.

[27] Trials in the two true belief conditions were defined in a somewhat different way in the preregistration, focusing on whether or the to-be-rated animal was the one linked with the first, the second-open or the closed box. We deviated from this grouping of the trials to keep the identity of the to-be-rated animal constant within experimental condition, just like it was on the underspecified belief trials, thus better match the TB and UB trials.

Just like in the previous experiments, after receiving the instructions for the main task, participants first performed four practice trials (two self- and two other-perspective trials, with one-one true and underspecified belief trial per perspective) without feedback. Following a reminder (regarding what counts as a valid trial and pointing out again what the agent believes about the turtle), they received 108 test trials, in two blocks: 18 underspecified, 18 true belief unambiguous, and 18 true belief ambiguous belief trials per perspective, with each alternative presented six times within each perspective and belief combination (i.e. the three animals rated an equal number of times within all three belief conditions).

The perspective, belief, the alternative type, the identity of the to-be-rated box was counterbalanced within blocks as well as the colour of the box that opened in the first step. The order of the trials was pseudorandomized within the blocks, such that there were no more than three consecutive trials with the same perspective, belief, alternative type, and to-be-rated animal in a row. Six different trial orders were used, the presentation of which was counterbalanced across participants.

As in the previous experiments, in addition to the experimental trials, participants were presented with 'attention checks trials', twice during the practice and twelve times during the test phase, to ensure that they pay attention to which box remained closed at the end. The structure of these trials was the same as in Experiment 3; their presentation was counterbalanced and randomized, such that participants could not predict when they would have to indicate the location of the hidden animal, based on the perspective prompt, the belief or the identity of the to-be-rated animal.

4.5.1.4. Data analysis

The data was analysed using the same dependent measures, data exclusion and participant inclusion criteria as before, with the threshold for the minimum number of correct responses on the attention check trials set at 66.66% (for the trial exclusion rates see: Supplementary Materials, **Table S4.2** and **S4.4)**. Lower than 0.75 mean rating on the SELF TB_UNAMB actual trials or higher than 0.25 mean rating on the SELF TB_UNAMB impossible1 trials (that is when the red box remained closed and the to-be-rated animal was the fox or chick, respectively) was taken as signs of misunderstanding the task or the use of the scale. Using these criteria resulted in the exclusion of N=1 participant.

Our crucial tests were: comparison of the ratings provided for 1) the actual and impossible1 alternative, on the SELF TB_UNAMB trials (i.e. for the fox and the chick, when the red box remained closed); 2) for the actual and possible alternative on the OTHER underspecified belief trials (i.e. for the

chick and the turtle when the yellow box remained closed); 3) for the possible and impossible alternative on the SELF underspecified and 4) for the impossible1 and impossible2 alternative on the SELF TB UNAMB trials (i.e. for the turtle and the fox when the yellow and for the turtle and the chick when the red box remained closed) as well as comparisons of the ratings of 5) the actual alternative on the SELF UB and TB_UNAMB trials (i.e. for the chick when the yellow and the fox when the red box remained closed). While the first and the second comparison tested whether participants themselves perform the necessary computations (i.e. infer the identity of the hidden animal), and whether they are able to track what conclusions the other might draw on the basis of the prior belief she holds, *explicitly*; the third, fourth and fifth tested our main hypothesis, whether they do this *spontaneously*.

Mean reaction times were investigated by running 3 x 3 repeated-measures ANOVAs, separately for the self- and other-perspective trials, with belief (TB_UNAMB, TB_AMB, UB) and alternative type (actual, possible/impossible2, impossible/impossible1) as within-subject factors, and subsequent paired-samples t-tests on the log-transformed RT data, focusing on the SELF and OTHER possible-impossible comparison listed above, to capture potential egocentric and altercentric interference effects.

As in the previous experiments, we also computed the difference of the possible and impossible trials' mean ratings, on the SELF underspecified and the impossible2 and impossible1 trials' mean ratings on the SELF true belief unambiguous trials, for each participant, to investigate how many participants demonstrated the predicted effect. If the difference of the two scores (SELF UB possible-impossible difference - SELF TB_UNAMB impossible2-impossibe1 difference) was larger than 0.01, this was considered as tentative evidence for spontaneously representing both alternatives (and thereby the conclusion) of the other agent.

4.5.2 Results

5.2.1 Rating analyses

Other-perspective trials

To investigate whether participants take into account the other agent's incorrect belief and the resulting different conditional rule she applies, when they are required to track what conclusions the other may draw from the beliefs she holds, first we analysed the likelihood estimations they provided on the other-perspective trials.

Friedman tests indicated a significant difference between the three types of alternatives, on all three types of belief trials (TB_UNAMB: $\chi^2(2)=55.70$, $p< .001$, Kendall's $W=0.796$; TB_AMB: $\chi^2(2)=55.70$, $p< .001$, Kendall's $W=0.759$; UB: $\chi^2(2)=58.80$, $p< .001$, Kendall's $W=0.840$). Post hoc Wilcoxon Signed Rank tests revealed that, on the true belief trials, participants' ratings were significantly higher for the actual, than for the other two alternatives, that is for the animal linked with the box that remained closed at the end (TB_UNAMB – fox (actual)-turtle (impossible2): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; fox (actual)-chick (impossible1): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; TB_AMB – turtle (actual)-fox (impossible2): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; turtle (actual)-chick (impossible1): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$). This indicates that participants understood that, in those cases when the agent's prior knowledge matched their own, she could draw the same conclusion they could and could safely infer the identity of the hidden animal. Although ratings were somewhat higher for the chick than either for the turtle on the TB-UNAMB or for the fox on the TB_AMB trials (see **Figure 4.12**), as the difference between the two impossible alternatives was very small on both trialtypes, and significant only in one (TB_UNAMB: $Z=-2.49$, $p_{adj(3)}= .013$, $r=0.421$), but not in the other condition (TB_AMB: $Z=-1.24$, $p_{adj(3)}= .215$, $r=0.209$), it most likely reflects mere random variation in the data.

Crucially, on the underspecified belief trials, when the box that remained closed was the one about which the agent mistakenly believed that it can be the hiding place of two animals (thus contains either the chick or the turtle), participants' ratings were significantly higher for both the actual and the possible alternative than for the impossible one, i.e. for the chick and the turtle compared to the fox ($M_{chick}=0.75$, $SD_{chick}=0.19$ $M_{turtle}=0.62$, $SD_{turtle}=0.20$ versus $M_{fox}=0.05$ $SD_{fox}=0.08$; chick (actual) -fox (impossible): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; turtle (possible)-fox (impossible): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$), indicating that, when performing the evaluations from the agent's perspective, participants took into account her incorrect belief and the fact that she represents two alternatives as a result of this belief on those trials where the yellow box remains closed at the end. Importantly, despite the two alternatives were equally likely for the other agent, participants provided higher ratings for the animal that was actually hiding in the yellow box (i.e. for the chick), based on what they could infer from the presented evidence, than for the animal about which only the agent believed that it could hide in the box (i.e. for the turtle), suggesting a marked difficulty with inhibiting their own knowledge ($Z=-4.39$, $p_{adj(3)}< .001$, $r=0.742$).

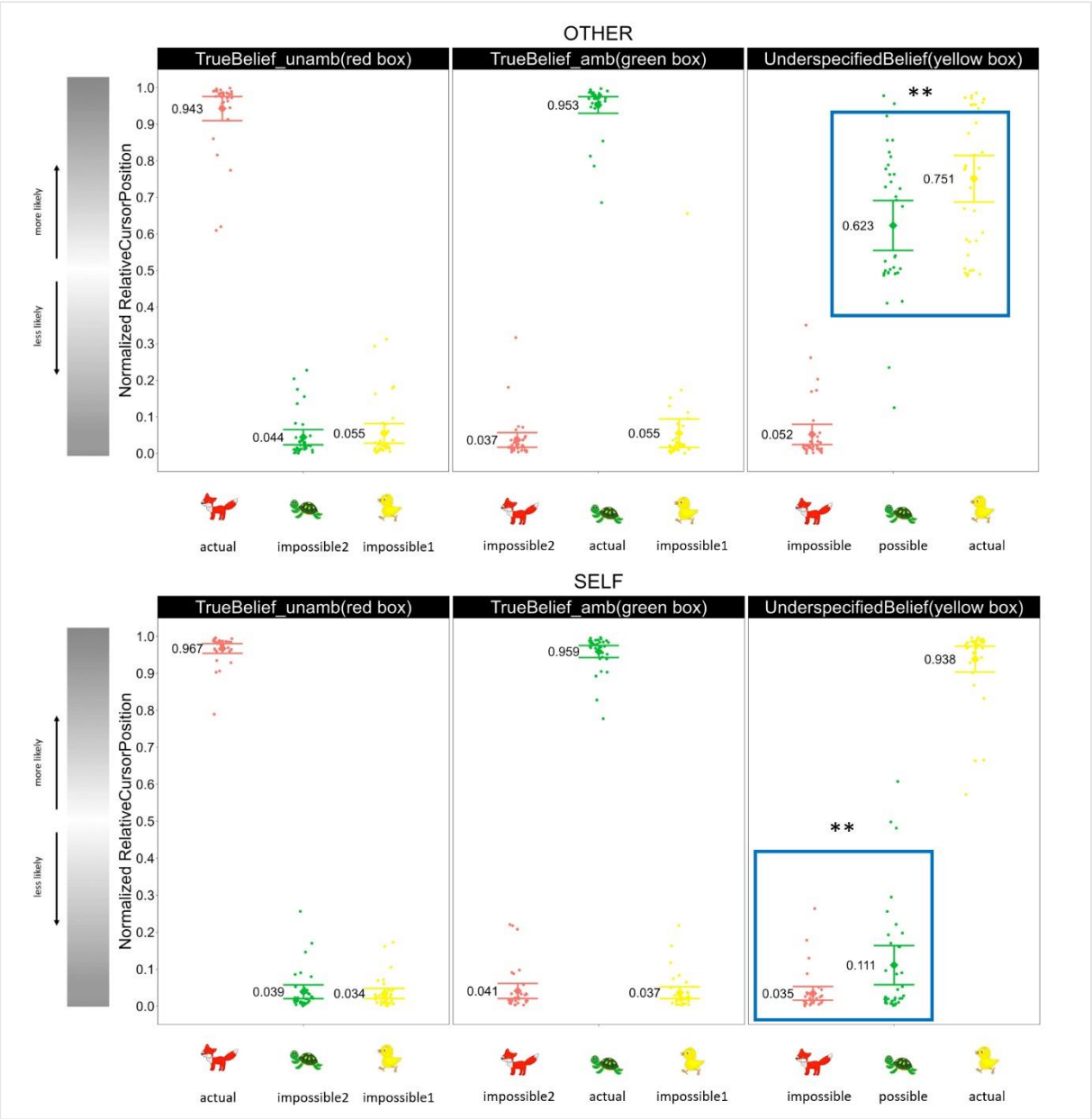


Figure 4.12 Mean ratings (mean normalized relative cursor position) on the self- and other perspective trials per belief and alternative type in Experiment 4. Belief was determined by which box remained closed at the end and alternative type by the to-be-rated animal's identity. Higher values indicate that participants considered it more likely that the presented animal was the one hidden in the scene, either from their own (SELF trials) or from the agent's perspective (OTHER trials). Error bars represent 95% CI, dots show the individual means. Blue frames indicate our two main foci of interest: the difference between the possible and impossible alternative on the self-perspective and the actual and possible alternative on the other-perspective underspecified belief trials. The first comparison investigates whether there is an altercentric bias (indicating a spontaneous representation of what the other considers possible) on the self-perspective trials, and the second test for the explicit representation of the two alternatives the agent could represent on the other-perspective trials). Stars indicate significant differences between the two experimental conditions. *: $p<0.05$, **: $p<0.01$

Self-perspective trials

To check whether participants themselves performed the necessary computations and, most importantly, to test for our main hypothesis, namely whether human adults track others' consecutive inferences (take into account the fact that they hold a different prior belief and therefore apply a different conditional rule than they do), *spontaneously*, next we analysed the ratings provided on the self-perspective trials.

Just like for the OTHER trials, Friedman tests indicated a significant overall difference between the three types of alternative on all three types of belief trials (TB_UNAMB: $\chi^2(2)=52.51$, $p< .001$, Kendall's $W=0.750$; TB_AMB: $\chi^2(2)=52.63$, $p< .001$, Kendall's $W=0.752$; UB: $\chi^2(2)=56.56$, $p< .001$, Kendall's $W=0.808$). Ratings were significantly higher for the actual than for the other two alternatives, on both the TB_UNAMB (fox (actual)-turtle (impossible2): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; fox (actual)-chick (impossible1): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$) and the TB_AMB trials (turtle (actual)-fox (impossible2): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; turtle (actual)-chick (impossible1): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$) as well as on the underspecified belief trials (chick (actual)-turtle (possible): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$; chick (actual)-fox (impossible): $Z=-5.16$, $p_{adj(3)}< .001$, $r=0.872$). There was no significant difference between the ratings provided for the two impossible alternatives on the true belief trials (TB_UNAMB: $Z=-0.07$, $p_{adj(3)}= .948$, $r=0.011$; TB_AMB: $Z=-0.00$, $p_{adj(3)}= 1.00$). Such results indicate that participants themselves had no problems with inferring the hidden animal's identity.

Crucially, on those trials where the yellow box remained closed at the end (underspecified belief trials), ratings were higher not only for the chick (actual alternative) but also for the turtle (the possible alternative, i.e. the animal only the agent linked with the box), compared to the fox, i.e. the impossible alternative ($M_{turtle}=0.11$, $SD_{turtle}=0.15$; $M_{fox}=0.03$, $SD_{fox}=0.05$; $Z=-3.58$, $p_{adj(3)}< .001$, $r=0.605$). This suggest that participants spontaneously applied their knowledge about the agent's incorrect belief (and its consequences), when they encountered with a situation where it was relevant, and represented both alternatives the agent ended up representing as a result of this belief. Contrary to our second prediction, there was, however, no significant difference between the underspecified and the true belief trials in terms of how participants rated the animals linked with the box that remained closed at the end, i.e. the actual alternatives (UB-TB_UNAMB: $Z=-0.93$, $p_{adj(2)}= .351$, $r=0.157$; UB-TB_AMB: $Z=-0.48$, $p_{adj(2)}= .635$, $r=0.081$). This indicates that participants either did not encode the fact that the agent assigned lower probability to the chick being in the yellow box than they did (as a result of representing two, equally likely alternatives) and/or this information had no effect on their own judgements. Taken together while we found evidence for a similar altercentric bias, we observed in Experiment 1 and 2, reflected by significantly higher ratings for the turtle, i.e. the possible alternative compared to the fox,

i.e. the impossible alternative, on the underspecified belief trial. However, our second prediction, that in this experiment such a bias will be also present in lower ratings for the actual alternative on these trials (compared to the ratings provided for the actual alternative on the true belief - unambiguous trials), was not supported by the data.

Regarding the individual data, the predicted effect (UB possible (turtle)-impossible (fox) difference score minus TB_UNAMB impossible2 (turtle)-impossible1 (chick) difference score > 0.01) was present in N=16 participants. Altogether N=24 participants demonstrated a difference score on the self UB and TB_UNAMB trials that was in line with our predictions (UB difference score > 0, after correcting for the TB_UNAMB difference). Excluding those participants who were not included in the RT analyses (N=2), did not change the main results (SELF TB_UNAMB actual-impossible1: Z=-5.12, $p_{adj(3)}$< .001, r=0.892; SELF UB possible-impossible: Z=-3.76, $p_{adj(3)}$< .001, r=0.655; SELF TB_UNAMB actual-UB actual: Z=-0.97, $p_{adj(2)}$= .660, r=0.873; OTHER UB possible-impossible: Z=-5.12, $p_{adj(2)}$< .001, r=0.892; OTHER UB actual-possible: Z=-4.17, $p_{adj(3)}$< .001, r=0.726).

5.2.2 Reaction time analyses

Other-perspective trials

To investigate whether the egocentric bias also emerges in participants' reaction times and to provide further evidence for the representation of both alternatives, on those trials where participants had to take the other agent's perspective, next we analysed participants' reaction times (i.e. the time necessary to perform their ratings) on the other-perspective trials. A 2 x 3 repeated-measures ANOVA, with belief (TB_UNAMB, TB_AMB, UB) and alternative type (actual, possible/impossible2, impossible/impossible1) as within-subject factors, revealed a significant main effect of belief (F(2, 64)=4.78, p= .012, η_p^2= .130), but no significant main effect of alternative type (F(2, 64)=0.69, p= .511, η_p^2= .021). There was, however, a significant belief x alternative type interaction (F(3.29, 105.39)=3.63, p= .013, η_p^2= .102), resulting from the fact that participants were the fastest to perform the likelihood estimations from the agent's perspective when the to-be-rated animal was the fox, independent of belief (see **Figure 4.13**), that is even when the animal that actually hid in the scene was the chick (UB trials) or the turtle (TB_AMB trials). Pairwise comparisons revealed that participants were significantly faster in providing their responses for the fox than for the other two animals on the TB_UNAMB (fox

(actual)-chick (impossible1): $t(32)=-2.78$, $p_{adj(3)}= .027$, $d=0.484$; fox (actual)-turtle (impossible2): $t(32)=-2.58$, $p_{adj(3)}= .030$, $d=0.449$) and marginally significantly faster on the TB_AMB trials (fox (impossible2)-turtle (actual): $t(32)=2.08$, $p_{adj(3)}= .096$, $d=0.362$; fox (impossible2)-chick (impossible1): $t(32)=-2.24$, $p_{adj(3)}= .096$, $d=0.390$). Though the difference was also present on the UB trials, it was not significant (fox (impossible)-chick (actual): $t(32)=-0.43$, $p_{adj(3)}= .627$, $d=0.074$; fox (impossible)-turtle (possible): $t(32)=-0.72$, $p_{adj(3)}= .474$, $d=0.126$). Relatedly, participants were significantly faster in providing their ratings for the actual alternative (the fox) on the TB_UNAMB than for the actual alternative (the turtle) on the TB_AMB ($t(32)=-4.01$, $p_{adj(2)}< .001$, $d=0.697$) and marginally faster than rating the actual alternative (the chick) on the UB trials ($t(32)=-1.92$, $p_{adj(2)}= .064$, $d=0.334$). Such results most likely reflect the fact that, unlike for the chick or for the turtle, for the fox it was possible to perform a one-to-one mapping between the box (location) and the animal (identity), both from self- and from other-perspective, which made the decisions easier for this animal. Finally, despite it took longer for participants to perform the ratings from the agent's perspective on the underspecified belief trials when the to-be-rated animal was the one about which the agent mistakenly believed that it can also hide in the yellow box (the turtle, *M*=1764 ms, *SD*=471 ms), compared to when it was the animal that actually hid there (the chick, *M*=1700 ms, *SD*=445 ms), the difference between the two alternatives was not significant ($t(32)=-1.23$, $p_{adj(3)}= .227$, $d=0.215$). Although none of the pairwise comparisons was significant on the underspecified belief trials, the overall pattern of results corroborates the findings obtained in the rating analyses, that participants found it difficult to inhibit their own knowledge (and make judgements according to that of the other) when the two conclusions did not match.

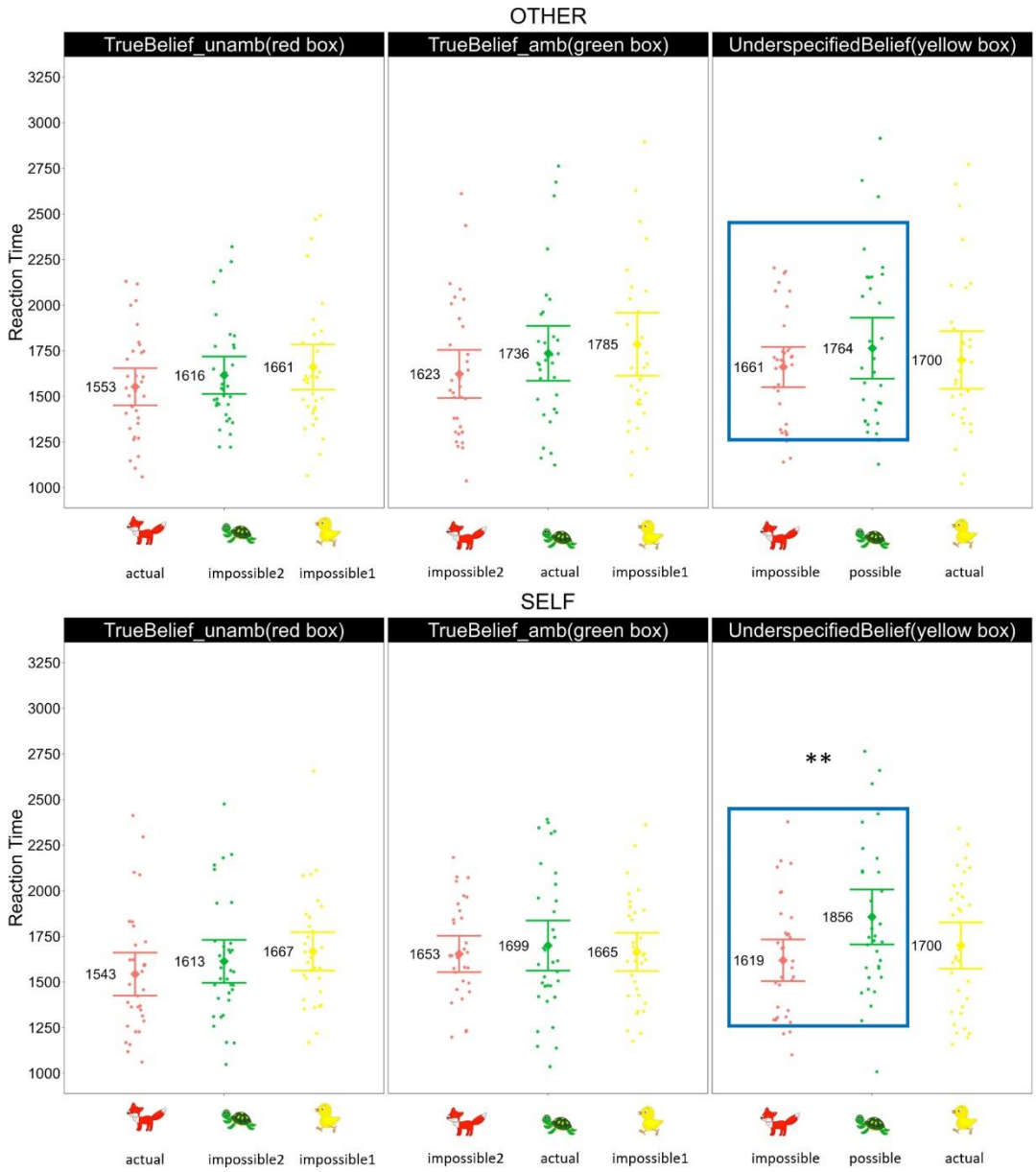


Figure 4.13. Mean reaction times (time necessary to perform the likelihood estimations) for the three types of alternatives on the self- and other-perspective true belief-unambiguous, true belief-ambiguous and underspecified belief trials in Experiment 4. Belief type was determined by which box remained closed at the end. Alternative type was determined by the to-be-rated animal's identity. Error bars represent 95% CI, dots show the individual means. Blue frames indicate our two main foci of interest: the difference between mean rating time of the possible and impossible alternative on the self- and on the other-perspective underspecified belief trials. Stars indicate significant differences between the two experimental conditions. *: p<0.05, **: p<0.01

Self-perspective trials

Finally, to investigate whether the agent's conclusion interferes with participants' own decision-making process, in case there is a mismatch, i.e. whether there are further signs of spontaneous tracking of the other agent's inferences, we performed the same analyses as above, on the self-perspective trials. A 2 x 3 repeated-measures ANOVA, with belief (TB_UNAMB, TB_AMB, UB) and alternative type (actual, possible/impossible2, impossible/impossible1) as within-subject factors, yielded a significant main effect of alternative type ($F(2, 64)= 3.27$, $p= .044$, $\eta_p^2= .093$) and belief ($F(2, 64)=11.66$, $p< .001$, $\eta_p^2= .267$) as well as a significant belief x alternative type interaction ($F(4, 128)=5.57$, $p< .001$, $\eta_p^2= .148$), reflecting the markedly different reaction time pattern on the three types of belief trials. While there was no significant difference between the three types of alternative in terms of reaction times on the TB_AMB trials i.e. when the green box remained closed at the end (all *t*s <0.540, all p_{unadj}s>.593), participants were faster to perform their estimations for the actually hidden than for the other two animals on the TB_UNAMB trials, with the difference being significant for the fox (actual) - chick (impossible1) ($t(32)=-3.41$, $p_{adj(3)}= .006$, d=0.593) and marginally significant for the fox (actual)-turtle (impossible2) comparison ($t(32)=-2.16$, $p_{adj(3)}= .076$, d=0.376). With respect to the two impossible alternatives, there was no significant difference in how fast participants rated those ($t(32)=-1.17$, $p_{adj(3)}= .253$, d=0.203). Altogether, these results suggest that participants could easily infer the identity of the hidden animal, particularly in those cases when it was possible to perform a one-to-one mapping between the box and the animal, both from self- and other-perspective.

Crucially, on the UB trials, i.e. when the agent represented two alternatives regarding the identity of the hidden animal, it took participants longer to provide their responses for the turtle, i.e. for the animal about which the agent mistakenly believed that it can also hide in the yellow box (possible alternative), than for either of the other two alternatives (M_{turtle}=1856 ms, SD_{turtle}=426 ms versus M_{fox}=1619 ms SD_{fox}=23 ms; M_{chick}=1700 ms, SD_{chick}=358 ms; turtle (possible) -fox (impossible): $t(32)=-4.50$, $p_{adj(3)}< .001$, d=0.783; turtle (possible)-chick (actual): $t(32)=-2.64$, $p_{adj(3)}= .026$, d=0.459), with no difference between the mean reaction times on actual and the impossible trials ($t(32)=-1.39$, $p_{adj(3)}= .176$, d=0.241), indicating a strong altercentric interference from the other agent's perspective. The time necessary to perform the estimations for the turtle (the possible alternative) was higher on these than either on the TB_UNAMB ($t(32)=-4.66$, $p_{adj(2)}< .001$, d=0.811) or on the TB_AMB trials ($t(32)=-2.81$, $p_{adj(2)}= p_{adj(2)}= .008$, d=0.490).

4.5.3 Discussion

The results of Experiment 4 suggest that human adults take into account another agent's incorrect belief and the different conditional rules she may apply as a result of the belief she holds, to represent what conclusions the agent may draw from a piece of information, not only when they have to do so but also when this would not be necessary for the task they perform. On those trials where the box that remained closed at the end was the one about which the agent incorrectly believed that it can serve as the hiding place of two animals (i.e. was the yellow), ratings were higher for the animal about which the agent mistakenly believed that it can hide in the box (the turtle) than for the animal about which neither the participants nor the agent believed that it can hide there (the fox), even when participants did not have to take the agent's perspective, indicating a spontaneous consideration of the conclusion the agent could draw in these cases. In addition, it took participants much longer to perform the estimations on these than on any of the other trials, implying a rather strong impact of the agent's belief on the decision process and/or difficulties to counteract its influence. Notably, the altercentric bias (i.e. higher ratings on the SELF underspecified 'possible' compared to the 'impossible trials), was not accompanied by lower ratings for the chick, i.e. for the animal that was actually associated with the yellow box, on the SELF underspecified belief trials. If participants had spontaneously represented not only the two alternatives the other agent did but also the fact that she assigns a 50% probability to each, this should have shifted the estimations not only for the turtle - from less towards 'more likely' - but also for the chick, in the other direction, on the underspecified belief trials. The fact that there was no sign of such an 'opposite' bias suggests that participants may have only represented the disjunction itself ('*it is either the chick or the turtle*') or fact that the turtle was a 'possible' alternative for the other agent, spontaneously, but not the probability of the two alternatives. Alternatively, it might happen that such a piece of information, even if encoded, can bias estimations for alternatives about which the other has a different belief (which was the turtle, in our case) and not for alternatives about which the other agent holds the same belief, specifically that the chick can go only to the yellow box.

Importantly, the observed effects (the altercentric bias and interference from the other agent's perspective) indicate that adults are able to track what inferences another agent may draw in a situation, spontaneously, even when those are made up of multiple steps (where the conclusion of one inference serves as the premise of another) and when correct attributions require the combination of different logical rules, if they do not have to constantly monitor what events the other has

witnessed, to be able to attribute the appropriate content, as they had to in Experiment 3. In the current experiment participants only had to pay attention to the colour of the box that remained closed at the end and apply their knowledge about the agent's prior belief (more specifically, what conditional rule she applies), in case it was the yellow, i.e. attribute an *a priori* computed belief content in case a certain 'trigger' was present. This clearly made the task easier than it was in Experiment 3: although we did not directly compare the two experiments, participants were visibly faster in this than in the previous, or actually in any of the previous experiments, on all but the SELF underspecified belief possible trials. The fast responses in turn might have contributed to the less precision and the more pronounced egocentric bias observed on the OTHER underspecified belief trials, via leaving little time for participants to inhibit the interfering irrelevant content, i.e. their own perspective. It is an intriguing question, and a matter of future research, whether our results are specific to the particular situation we investigated or the tracking of other agents' inferences is generally easier when it requires only the consideration of the others' prior assumptions but not the tracking of what information they acquire on a trial-by-trial basis and what are the limits of inference-tracking in such cases.

4.6. General Discussion

The purpose of the present study was to investigate human adults' ability to track the logical inferences of other agents, i.e. what conclusions others may draw from the beliefs they hold. Specifically, we tested whether adults represent what inferences another agent may make on the basis of the events she saw and/or her prior assumptions, in a given situation, regarding an object's identity or location, (i) when they are instructed to take the other's perspective and (ii) when they do not have to track what the other believes or knows. To address this question, in four experiments we presented participants with picture-sequences, depicting a girl and three boxes. Two of the boxes always opened one after another, providing visual access to their content, while the third always remained closed. The revealed evidence made it possible for participants to infer a hidden animal's location or identity, by applying disjunctive syllogism (Experiments 1 and 2) or by combining the output of a disjunctive syllogism with a previously acquired conditional rule (Experiments 3 and 4). We measured how likely participants consider that the animal is hiding at a certain location or that a certain animal has hidden in the scene, either from their own or from the other agent's perspective. The agent had either the same knowledge, hence could draw the same conclusion participants could, or had access to less

evidence (or held a different prior belief) than participants did and therefore arrived at a different conclusion, more specifically, ended up representing two, equally likely alternatives regarding the hidden animal's location/identity. We were interested in whether in this latter case participants take into account that the agent considers a box or an animal they could exclude from the range of options as a potential hiding place or a potentially hidden animal, (i) when they perform the likelihood ratings for this particular box or for this particular animal from the other agent's perspective, and (ii) when they perform the same judgements from self-perspective. In specific, we investigated whether this knowledge (what the other agent considers 'possible') biases their own likelihood estimations and interferes with the process of decision-making, in the same way the false belief of another agent does in other ToM tasks using continuous measures (see e.g. Marshall, Gollwitzer, & Santos, 2018; Speiger et al., 2021).

In all four experiments, we found strong evidence that human adults can and do take into account what another agent knows and doesn't know (based on what events she had visual access to or what she was told), hence what conclusions she may draw in a situation, when they have to track the other's beliefs *explicitly*. When participants had to estimate the likelihood that the target animal hid at a certain location / or that a certain animal has hidden in the scene from the agent's perspective, and the agent could not unambiguously infer the location/identity of the hidden animal (either as a result of not witnessing the second box opening or holding an incorrect prior belief), participants' ratings were similar for the actual location/actually hidden animal and for the one only the agent considered 'possible', indicating that they understood that the agent represented two, equally likely alternatives. In line with previous findings on adults' propensity to attribute their own knowledge to others in general (Birch & Bloom, 2007; Dumontheil et al., 2010; Keysar et al., 2003) and the impact of working memory load on their capacity to take others' perspective in particular (Cane et al., 2017; Lin et al., 2010; Qureshi et al., 2010), a clear egocentric bias also emerged on these trials, the magnitude of which scaled with the complexity of the inference participants themselves had to perform (though was also affect by the time participants allocated to the estimation).

Crucially, in three out of the four experiments (Experiments 1, 2 and 4), participants rated the possible alternative higher (compared to the impossible one) not only on those trials where they had to take the other agent's perspective, when the agent represented two alternatives, but also on those where this was unnecessary, as they only had to perform first-person judgements. These findings suggest that participants *spontaneously* represented the conclusions the other agent could draw from the beliefs (i.e. 'premises') she had. In two of the three experiments (Experiment 2 and Experiment 4) this altercentric bias was accompanied by longer rating times, providing further evidence for the spontaneous consideration of the other's (potential) belief content and, more generally, for the imperfect separation of the self- and the other-perspective. Importantly, these results show not only

that human adults track other agents' logical inferences but also that they spontaneously compute other agents' underspecified belief contents, for which we failed to find strong and convincing evidence in the study presented in the previous chapter, using another paradigm and different measures.

The current findings contribute to the research on adults' spontaneous ToM abilities in at least two ways. First, theoretically, by showing that these abilities extend over and above the computation and attribution of false beliefs, to (i) more complex ToM computations, that reflect a genuine understanding of the functional role of beliefs in 'mental economy' (Rakoczy, 2012), namely that beliefs yield other beliefs, and to (ii) the computation of underspecified belief contents, i.e. belief contents that do not make the other's behaviour fully predictable. Second, methodologically, by demonstrating that altercentric effects emerge not only in reaction times and categorical judgements of participants (for review see: Kampis & Southgate, 2020) but also in likelihood estimations of different events, at least if those are provided on a continuous rating scale. They also point out that using such 'probability scales' or, more generally, continuous measures may be a more fruitful approach to investigate whether and how human adults represent other agents' underspecified beliefs than the one we took in the previous chapter.

Notably, the results do not only demonstrate the scopes but also the limits of the ability as well as of the method we used. The altercentric effects were clearly absent in Experiment 3, indicating that adults may not track other agents' chains of inferences spontaneously, at least when this requires the constant monitoring of what information others have access to be able to perform the appropriate attributions. Such a finding is not surprising given that doing so, i.e. tracking inference chains without conscious control over the process, could easily lead to prediction errors (confusing the self- and the other-perspective), hence it may not be adaptive from an evolutionary perspective. In addition, despite the predicted estimation bias emerged in all but one experiment, it was very small, and clearly present in only 25-40% of the sample when the task required continuous tracking of what events the other witnessed, with its magnitude varying to a large extent even in those who demonstrated the effect. One potential reason for this is that each experimental condition consisted of only six trials and, if the bias did not consistently appear each and every time participants had to rate the alternative that was 'possible' for the other, this trial number might not have been enough to capture the impact of the represented belief content on participants own judgements. Another option is that our scale had clear endpoints, and even though these were not labelled, the presence of such 'anchors' may have made it easier for participants to counteract the effect of the other's belief content. Both of these possibilities should be investigated in the future, by presenting more trials (that would also make it possible to test for potential order effects that may as well exist) and by using other types of continuous measures.

Throughout the whole study, we assumed that belief attribution took place in a prospective manner: those who did track what conclusions the other agent may draw from the beliefs she had, computed her belief content either (i) at each step of the unfolding inferential process, ascribing even the intermediate conclusions to the other (ii) when it became clear that the two conclusions would differ (e.g. when the agent lost track of the events) (iii) and/or when participants drew the final conclusion from first-person perspective. Yet, based on the current results, we cannot exclude the possibility that the content of the agent's belief was computed in a retrospective manner, when participants received the test question, which prompted them to evaluate their own belief (and possibly their own (un)certainty in the conclusion they have drawn from the premises). The test question may have acted as a trigger for recalling the previous events and, along with this, what information the other agent has, resulting in performing the conclusion again, from her perspective. Alternatively, the very same test question might have activated the representation of the already computed belief content. Whether altercentric effects observed in the present study emerged due to the presence of this 'trigger', as the result of the fact that, in this study, on half of the trials participants had to take the other's perspective (which might have drawn attention to the agent's beliefs as well as to the difference between the two perspectives), or because of the probabilistic nature of the measure we used to capture the effect, is a matter of future research. Further studies should assess whether the effect emerges: 1) if the agent is completely irrelevant and 2) when other types of measures are used as well as 3) when exactly the agent's belief content is computed during the task (by measuring, for instance the time necessary to perform the different inferential steps or manipulating the time limit available at each step). Finally, besides investigating what are the necessary and sufficient preconditions for the emergence of the effect, i.e. for the spontaneous tracking of others' inferences, at group level, future studies should also address what individual differences determine whether or not someone tends to spontaneously engage in such a ToM process.

Chapter 5: General Discussion

5.1 Summary of the findings

Successful navigation in the social world requires correct prediction and interpretation of other agents' actions. It has been widely accepted that to do so humans rely on their ability to represent others' mental states, such as their goals, beliefs and desires, and take into account that those may differ from what they want, believe or know. This capacity, usually referred to as theory of mind (ToM), has been the target of extensive research in the past few decades, with studies predominantly focusing on situations where the other is mistaken about the actual state of affairs, for example, an object's location, and participants have to predict how the agent will act, on the basis of the false belief the other holds (Wellman, 2001). In the last two decades growing amount of evidence suggests that human beings compute the content of other agents' false beliefs and visual perspectives spontaneously, that is even when this would not be necessary in the given situation (Schneider et al., 2017; Kampis & Southgate, 2020), and do this from very early on (Scott & Baillargeon, 2017), as indicated by their performance on nonverbal versions of the classic ToM tasks (although some of these results were not replicated recently, see: Burnside et al., 2018; Kulke, von Duhn, et al., 2018). Some authors argue that these results reflect the operating of an early emerging ToM-like system that is much restricted in terms of what kind of information it can encode, compared to late-emerging fully-fledged ToM, claiming that ToM cannot be efficient and flexible at the same time (Butterfill & Apperly, 2013; Rakoczy, 2017). Others claim that they reflect a genuine understanding of the other agent's mind (Baillargeon et al. 2018; Carruthers, 2017) and it is exactly this propensity to spontaneously engage in mentalizing what enables the smooth unfolding of social interactions in the everyday life of humans (see e.g. Kovács, 2016). Despite the fact that more and more results seem to support this latter view, specifically the idea that verbal and nonverbal ToM tasks are subserved by the same ToM mechanism, the field is still dominated by the debate between the two approaches and by studies investigating the attribution of false beliefs about various object properties, with the aim to settle it by providing evidence in favour of one approach or the other.

After reviewing the findings in the literature and the ongoing debate on the nature of ToM, in Chapter 1, we have argued that people encounter a much wider range of social situations than the ones that have been extensively investigated so far, in which the smooth adaptation to the other's behaviour requires participants to perform different ToM computations or to compute different kinds of mental states that have been the target of research up until now. The present thesis focuses on three such computations, that likely play an important role in the flexible adaptation to other agents' behaviour: 1) updating other agents' mental states based on the behaviour they demonstrate in a situation (assessed in Study 1); 2) tracking the hypotheses other agents entertain (investigated in Study 2); and

3) representing the conclusions others may draw from the beliefs they hold (addressed in Study 3). More specifically, we tested whether human adults perform these computations spontaneously, without having the intention to do so, using 'unintentional' as the defining criterium of 'spontaneous'.

In Study 1 we found that human adults update the mental state of another agent and revise their expectations regarding the agent's future behaviour if they observe the other repeatedly acting in a way that is incompatible with their original assumptions about the belief she/he holds, specifically with their assumptions about how the other encodes a certain colour, even if this is not necessary for the task they perform. This was indicated both by the anticipatory looking behaviour and the reaction time patterns participants showed. Signatures of updating emerged only after a few observations of the actions violating participants' expectations, both when correct anticipation required overcoming an explicit categorization rule (Experiment 1, i.e. 'put the green into the green box') and when there was no such rule (Experiment 2), the partner's actions rather reflected her subjective evaluation of the colours. Such results corroborate and extend recent findings which demonstrated that even very young children update previously attributed beliefs spontaneously, recomputing the content of the other's belief in a retrospective manner and acting accordingly, if they are provided with a new piece of information which suggests that the other might have witnessed certain events before (Király et al., 2018). Importantly, in both experiments of Study 1, roughly half of the participants (members of the 'update' subgroup) provided clear evidence that changes in their anticipatory looking behaviour reflected the updating of the other's mental state content and not that of a nonmentalistic rule. In particular, when later they were explicitly asked to take the other's perspective, in a categorization task, these participants could categorize items from the other's (updated) point of view, suggesting that for them the output of the process recruited in the anticipatory task was consciously accessible for later use. The rest of the participants (members of the 'noupdate' subgroup) did not provide such clear evidence for updating the other's mental state, i.e. they did not take the other's updated perspective in this later explicit perspective-taking task. Yet, this does not mean that they did not engage in mentalizing at all. In fact, the pattern of results, namely the fact that signatures of updating emerged by the end of the anticipatory task and the other's perspective seemingly interfered with participants' own decision making even in this subgroup, in the explicit perspective-taking task of Experiment 2, indicated that, they may have simply struggled to find a satisfying interpretation for the change in the other's behaviour, within the available time frame. Importantly, in addition to being able to categorize explicitly from the other agent's perspective, 'update' participants also took into account what they have learnt about the other's perspective, spontaneously, in an interactive task immediately following the one that prompted belief revision, where taking the other's perspective could be considered useful. The extent to which they did so was, however, surprisingly low, which raises

important questions regarding the relationship between the capacity to update the content of another agent's mental state and the propensity to spontaneously generalize the newly acquired knowledge about the other's stable beliefs (i.e. 'he thinks this ambiguous green colour is blue') to subsequent interactions. Future studies should address both this relationship and the factors determining the ease with which people update the content of other agents' mental states on the basis of their behaviour.

While in Study 1 we examined a ToM capacity that plays a crucial role in whether or not humans can smoothly adapt to unexpected changes in the other's behaviour *in the present*, in Study 2 and Study 3 we investigated ToM abilities that may play an important role in whether people can react quickly to the other's behaviour, *if this becomes necessary in the future*. In both studies, we focused on a situation that is quite common in everyday life, yet so far it has been unexplored, specifically when another agent is uncertain about the actual state of affairs, and hence must represent multiple, equally likely alternatives, simultaneously. We use the term 'underspecified' to refer to the belief the other agent holds in such situations, to emphasize that these beliefs, despite restricting the range of actions one can expect from the other agent, do not make the other's behaviour as predictable, as simple true or false beliefs do.

In Study 2 we aimed to test whether human adults represent the alternatives another agent represents regarding the location of an object, spontaneously, via measuring whether they allocate more attention to, and consequently, detect irrelevant changes in a dot's colour better at the locations corresponding to the agent's hypotheses. The study yielded rather inconclusive results. We found that when participants were explicitly instructed to track the agent's belief (Experiment 2), and she represented two, equally likely alternatives regarding where the object may be, due to not witnessing the hiding, participants were better in detecting changes not only at the actual location of the object but also at the location which was empty but could be considered a potential hiding place by the agent. That is, the voluntarily computed underspecified belief content of the other agent influenced their performance even in a task in which it was completely irrelevant, corroborating previous findings, which indicate that once computed, the content of other agents' belief influences human adults' task-performance in an uncontrollable manner, i.e. whether or not it has relevance for the actual task of the observer (Hegedűs & Király, 2022; Kovács et al., 2010; Samson et al., 2010; Schneider, Bayliss, et al., 2012a). Crucially, however, after an initial positive finding (Experiment 1a), we failed to find strong and consistent evidence for the spontaneous representation of the other agent's hypotheses. In those experiments where participants were not required to track what the other agent considers 'possible', the expected attentional bias (i.e. faster detection of the irrelevant change at the location the agent considered a potential hiding place) did not emerge, at least in the conditions where the object was present (Experiment 1b, Experiment 3a and 'ball present trials of Experiment 3b). This was the case,

despite participants seemed to register the difference between their own knowledge state and that of the agent, as indicated by the different pattern of miss ratios in the condition when the agent held an underspecified (as opposed to a true) belief. These results may indicate that human adults do not spontaneously compute the content of other agents' underspecified beliefs, either because of their complexity or because they do not provide a strong basis for action prediction. However, the general pattern of findings, and especially the results of our final experiment, in which the predicted effect emerged after eliminating a possible 'reality bias' by removing the object ('ball absent' trials of Experiment 3b), suggest that specific features of our methodology might have been responsible for the observed null results. In particular, the measure we have used (change detection sensitivity at a particular location to measure a specific bias in participants' spatial attention) might not have been sensitive enough to capture the potentially small impact of the other's belief content on participants' own representation of the state of affairs, due to other factors simultaneously influencing the allocation of spatial attention, such as the well-documented 'pull-of-the-real' (see e.g. Schneider, Lam, et al., 2012a for a similarly strong reality bias in anticipatory looking measures). The results of Study 3 corroborated this latter interpretation.

In Study 3 we investigated human adults' ability to represent what conclusions another agent may draw from the beliefs she holds, when the agent's inferential process results in the representation of one or more alternatives. We used situations in which the location or the identity of an animal could be inferred by applying disjunctive syllogism (Experiment 1 and 2) or by combining disjunctive inference with conditional reasoning (Experiment 3 and 4) both from first-person perspective, and via adopting the perspective of another agent, who observed the events. Specifically, we tested whether participants take into account that in certain situations the agent ends up representing two alternatives (and not just the one they themselves do) when they estimate the likelihood of different alternatives from the agent's perspective and, most importantly, whether an 'altercentric' estimation bias emerges when they perform the ratings from self-perspective (indicating that they also consider this *spontaneously*). Our results showed that participants represented the agent's alternatives, not only in the first case, when they had to take the other agent's perspective, but also in the second, when this was not necessary in the given situation. The magnitude of the observed altercentric effect was small, but present in three different experiments (Experiments 1, 2 and 4). However, it disappeared when the complexity of the to-be-represented content and of the computations participants had to perform (in order to attribute the appropriate content to the other agent) was higher (Experiment 3) than in the experiments before. Such results provide evidence not only for the spontaneous tracking of other agents' logical inferences but also for the spontaneous representation of the content of the other agent's underspecified belief, for which we failed to find convincing evidence in Study 2. While

the different results in Study 2 and Study 3 may simply stem from the fact that likelihood estimations and/or responses provided on a continuous scale are more sensitive to the content of another agent's belief, more specifically, what another agent considers 'possible', than other measures, it is important to note that they may also reflect genuine differences in the ToM computations participants performed in the two studies. In particular, it might happen that participants in Study 3 computed the content of the other agent's underspecified belief on the self-perspective trials only because self and other trials were intermixed within blocks, which may have made the agent's belief more relevant for the task, in general, and would not have done so if the other agent's underspecified belief had been completely irrelevant, as it was in Study 2. On this account, one may argue that Study 3 stands halfway through the semi-explicit and the fully implicit experiments of Study 2, as in Study 3 the other's belief was not completely task-irrelevant, yet its content was not deliberately tracked by the participants, on the self-perspective trials. Note that, however, in Study 3 participants knew whether they would have to perform the estimation from their own or from the other agent's perspective from the very beginning of the trials, therefore, they could have simply ignored the other's belief on the 'self' trials. Nevertheless, they did not, which supports the spontaneous nature of the computations underlying the observed effects. Future studies should investigate whether the estimation bias, present in three experiments of in Study 3, emerges if participants do have to take the other agent's perspective on any of the trials, i.e. whether people spontaneously represent the conclusions another agent may draw if the other's belief is fully irrelevant for the task (as it was in Study 2). Further studies should address the research question using other implicit measures that do not require participants to evaluate their own beliefs.

5.2 The functioning of adult theory of mind

What do our findings tell us about ToM functioning, in general? Considering that in Study 1 and Study 3 participants seemed to engage in fairly complex mental state reasoning without being instructed to do so, spontaneously performing computations, the complexity of which extends way beyond the scopes of minimal mindreading system, proposed by the two-system account of ToM (see: Butterfill & Apperly, 2013), our results provide support for the claim that ToM can operate flexibly yet, at the same time, efficiently in human adults. Therefore, there may be no need to postulate the existence of two distinct 'theories of mind' to explain human beings' success in everyday social life. Importantly, however, 'efficient' operation does not mean a fully 'automatic' functioning, at least not in those

situations we investigated in our studies. The results of Study 1, in specific, the fact that (i) signatures of updating emerged earlier when we decreased the inhibitory demands of the task, and that (ii) a number of participants could not find a satisfying explanation for the other's behaviour (and adjust to the change) within the available time frame even in this situation, indicate, for instance, that whether or not people update a previously attributed mental state when they observe a behaviour that does not correspond to their original assumptions, may depend heavily on the availability of executive resources. In a similar vein, considering the results of Study 2 and Study 3, in particular, the fact that consistent evidence for the spontaneous representation of the other agent's underspecified belief emerged only when the other's belief was, to some extent, relevant for the participants' task, suggests that while people may encode the content of other agents' hypothesis space, without having the intention (and being aware) of doing so, the process may not be purely stimulus-driven (i.e. triggered by the mere presence of another agent). Based on these findings, one could argue that the results of Study 1 are the product of a conscious, voluntary effort to interpret the observed behaviour, especially given the fact that, for some participants, the output of the update process was accessible, both for later use and for verbal report. Similar reasoning could be applied to Study 3. While we cannot completely exclude the possibility that the updating of the other agent's mental state in Study 1 was the result of an explicit mental state reasoning, in case of the 'update' participants, it is important to note that the fact that the information was accessible later, upon providing a prompt in the explicit perspective-taking task, does not mean that the process itself was conscious at any timepoint before receiving that prompt. Nor does it mean that participants intentionally engaged in mentalizing during the anticipatory task of Study 1. Notably, there was no sign of conscious, deliberate tracking of the other agent's belief in Study 3. In fact, according to the feedback provided at the end of the experiments, the great majority of the participants could not even figure out the general purpose of the study (only 1 to 3 out of the 36 participants in the different experiments mentioned that it was testing their ability to consider the perspective of the other). Here we argue that instead of reflecting the operation of 'the explicit theory-of-mind' our, seemingly contradictory findings across the three studies, merely provide further examples that the specific features characteristic of automatic and controlled processing can co-occur in adults' ToM.

So far, we have focused predominantly on our group-level findings and what these may teach us about how ToM operates in adults. Crucially, however, we did not only test for the predicted effects at group-level, but we also investigated how many participants demonstrated a performance pattern that was in line with our hypothesis, in each and every experiment. To much of our surprise, in all three studies, we observed a rather high individual heterogeneity both in terms of the presence and the magnitude of the predicted effect. In Study 1 we had clear evidence for the updating of the other's mental state in only 40-50% of the participants, based on their performance on the explicit perspective-taking task

we administered at the end of the experiments. The numbers were even more striking in Study 2 and Study 3: even in those experiments which yielded positive findings at group-level, only 20-50% of the participants demonstrated the predicted effect. In some cases, the difference between the critical conditions was driven by only a few participants. One may argue that this heterogeneity is simply the result of a natural fluctuation in participants' performance, which one should consider an experimental 'noise', hence it should be ignored. If we accept, however, that the individual scores we created reflect genuine differences in whether or not participants have engaged in the particular ToM computation investigated in the given experiment, then these results are quite puzzling. Is this huge variability a feature of all studies targeting spontaneous ToM in adults, or only ours, which investigated the ability to spontaneously perform relatively complex ToM computations? The answer is currently unknown, given that it seems rather uncommon to report data on how many participants demonstrated a performance indicative of spontaneous belief tracking (or the spontaneous computation of the content of the other agent's visual perspective) in the given study, at least in the great majority of the studies we are aware of. This may have several reasons, including a simple methodological one, namely that it is quite difficult to establish what can be considered good enough evidence for the spontaneous representation of the other agent's mental state content, at the level of the individual – something we also struggled with within our studies. One exception is a study by Bukowski and Samson (2017), which analysed the data from six dot perspective-taking studies, to identify the cognitive dimensions underlying individual differences in Level-1 perspective-taking. The authors found that adults vary to a large extent in terms of how well they handle conflicting perspectives and how much they prioritize their own versus the other's perspective (operationalized as the difference between consistent and inconsistent trials' and the self and other trials' inverse efficiency score, respectively). Although they did not directly investigate how many participants demonstrated the 'altercentric effect', the fact that only about 20% of the sample could be classified as 'altercentric' (someone focusing more on the other's than his/her own perspective) implies that the ratio of those who actually tracked the avatar's visual perspective, spontaneously, might have been close to what we observed in our experiments. Such a high heterogeneity led the authors to conclude that the ability to spontaneously track the mental states of others may "not be as universal as previously thought" (Bukowski and Samson, 2017, p. 11).

While we agree that individual differences in the actual *ability* to quickly compute and take into account the content of others' perspectives may indeed play some role in the marked inter-individual variability observed in our experiments (and assumed to be also present in many others, targeting spontaneous ToM), here we would like to propose another, slightly different interpretation. Namely, we argue that the heterogeneity may also result from huge individual differences in *when* (under what circumstances) people engage in more complex forms (or perhaps in any kind) of spontaneous

mentalizing, with the difference being not in *whether one is able to* perform such computations, but in whether he/she does so in a certain situation Human adults *can* readily update other agent's mental states (upon observing a behaviour that is incompatible with one's original assumptions), spontaneously track other agents' logical inferences and (at least under some circumstances) represent the alternatives others do, as indicated both by group and individual-level results. In our view, these abilities are part of the social-cognitive repertoire they are endowed with, to be able to function efficiently in the social world. Nevertheless, they often *do not* spontaneously perform these computations (or are far from being efficient in doing so), under the circumstances we investigated those, that is in non-interactive contexts, where the agent's belief has little or no relevance for their own goals. We argue, that instead of reflecting the limitations of spontaneous ToM capacities, such a difference between what people *can* do and what they actually *do*, rather highlights the flexibility with which ToM operates in human adults. It does not automatically 'turn on' in each and every situation where another agent is present, nor it may have a universal threshold, in the form of a set of necessary and sufficient preconditions, above which it becomes 'active' in everyone. In line with a recent proposal by Westra (2017), who, aiming to explain the contradictory findings in the spontaneous perspective-taking literature, argued that human adults deploy implicit ToM capacities in a context-sensitive, cost-efficient manner, we suggest that whether or not people engage in certain, complex forms of mentalizing in a particular situation depends on the actual goals and needs they have. More precisely, it depends on whether the cognitive system evaluates the information the given ToM computation can yield as something *possibly useful*, worth investing the effort in the given context. This conceptualization of how ToM operates in human adults may not only explain the huge variation in our data but may also advance our understanding of the large inter- and intra-individual differences one can observe in the everyday social functioning of humans.

Future studies should address how widespread the large individual variation in adults' performance on spontaneous ToM tasks is. If this phenomenon is indeed prevalent across studies, they should examine its underlying causes and provide further, more direct evidence for the proposed 'utility-based' operation of implicit ToM. Finally, further studies should investigate how the flexibility with which people deploy their spontaneous ToM capacities develops in ontogeny and how it relates to other cognitive abilities, such as the individual's executive functions.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38(4),* 419-439. https://doi.org/10.1006/jmla.1997.2558

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502-518. https://doi.org/10.1016/j.jml.2006.12.004

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

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic?. *Psychological Science*, *17*(10), 841-844. https://doi.org/10.1111/j.1467-9280.2006.01791.x

Back, E., & Apperly, I. A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, *115*(1), 54-70. https://doi.org/10.1016/j.cognition.2009.11.008

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, *46*, 112-124. https://doi.org/10.1016/j.cogdev.2018.06.001

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349. https://doi.org/10.1016/j.cognition.2009.07.005

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1-10. https://doi.org/10.1038/s41562-017-0064

Bardi, L., Desmet, C., Nijhof, A., Wiersema, J. R., & Brass, M. (2017). Brain activation for spontaneous and explicit false belief tasks overlaps: new fMRI evidence on belief processing and violation of expectation. *Social Cognitive and Affective Neuroscience*, *12*(3), 391-400. https://doi.org/10.1093/scan/nsw143

Bargh, J. A. (1994). The four horsemen of automaticity: Intention, awareness, efficiency, and control as separate issues. In R. Wyer & T. Srull (eds.), *Handbook of Social Cognition.* Lawrence Erlbaum.

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

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407-418. https://doi.org/10.1023/a:1023035012436

Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International review of research in mental retardation*, *23*, 169-184. https://doi.org/10.1016/S0074-7750(00)80010-5

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, *57*, 101350. https://doi.org/10.1016/j.infbeh.2019.101350

Birch, S. A. J., & Bloom, P. (2003). Children Are Cursed: An Asymmetric Bias in Mental-State Attribution. *Psychological Science*, *14*(3), 283–286. https://doi.org/10.1111/1467-9280.03436

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382-386. https://doi.org/10.1111/j.1467-9280.2007.01909.x

Braine, M. D., O'Brien, D. P., Noveck, I. A., Samuels, M. C., Lea, R. B., Fisch, S. M., & Yang, Y. (1995). Predicting intermediate and multiple conclusions in propositional logic inference problems: Further evidence for a mental logic. *Journal of Experimental Psychology: General*, *124*(3), 263-292. https://doi.org/10.1037/0096-3445.124.3.263

Bukowski, H., & Samson, D. (2017). New insights into inter-individual variability in perspective-taking. *Vision, 1*(*1*), 8. https://doi.org/10.3390/vision1010008

Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*, *46*, 4-11. https://doi.org/10.1016/j.cogdev.2017.08.006

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337-342. https://doi.org/10.1016/j.cognition.2009.05.006

Buttelmann, F., & Buttelmann, D. (2017). The influence of a bystander agent's beliefs on children's and adults' decision-making process. *Journal of Experimental Child Psychology*, *153*, 126-139. https://doi.org/10.1016/j.jecp.2016.09.006

Buttelmann, F., & Kovács, Á. M. (2019). 14-Month-olds anticipate others' actions based on their belief about an object's identity. *Infancy*, *24*(5), 738-751. https://doi.org/10.1111/infa.12303

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606-637. https://doi.org/10.1111/mila.12036

Cane, J. E., Ferguson, H. J., & Apperly, I. A. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 591. https://doi.org/10.1037/xlm0000345

Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology*, *20*(3), 393-420. https://doi.org/10.1348/026151002320620316

Carruthers, P. (2017). Mindreading in adults: Evaluating two-systems views. *Synthese*, *194*(3), 673-688. https://doi.org/10.1007/s11229-015-0792-3

Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, *11*(1), 1-9. https://doi.org/10.1038/s41467-020-19734-5

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, *359*(6381), 1263-1266. https://doi.org/10.1126/science.aao3539

Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*, 48-64.https://doi.org/10.1016/j.newideapsych.2015.01.003

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

Cohen, A. S., & German, T. C. (2009). Encoding of others' beliefs without overt instruction. *Cognition*, *111*(3), 356-363. https://doi.org/10.1016/j.cognition.2009.03.004

Conway, J. R., Lee, D., Ojaghi, M., Catmur, C., & Bird, G. (2017). Submentalizing or mentalizing in a Level 1 perspective-taking task: A cloak and goggles test. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 454. https://doi.org/10.1037/xhp0000319

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294-300. https://doi.org/10.1016/j.tics.2006.05.004

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12-30. https://doi.org/10.1016/j.cogdev.2018.01.001

Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*(2), 331-338. https://doi.org/10.1111/j.1467-7687.2009.00888.x

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, *21*(4), 419-460. https://doi.org/10.1016/S0364-0213(99)80029-9

El Kaddouri, R., Bardi, L., De Bremaeker, D., Brass, M., & Wiersema, J. R. (2020). Measuring spontaneous mentalizing with a ball detection task: putting the attention-check hypothesis by Phillips and colleagues (2015) to the test. *Psychological Research*, *84*(6), 1749-1757. https://doi.org/10.1007/s00426-019-01181-7

Elekes, F., Varga, M., & Király, I. (2016). Evidence for spontaneous level-2 perspective-taking in adults. *Consciousness and Cognition*, *41*, 93-103. https://doi.org/10.1016/j.concog.2016.02.010

Elekes, F., Varga, M., & Király, I. (2017). Level-2 perspectives computed quickly and spontaneously: Evidence from eight-to 9.5-year-old children. *British Journal of Developmental Psychology*, *35*(4), 609-622. https://doi.org/10.1111/bjdp.12201

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective-taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327-339. https://doi.org/10.1037/0022-3514.87.3.327

Fizke, E., Butterfill, S., van de Loo, L., Reindl, E., & Rakoczy, H. (2017). Are there signature limits in early theory of mind? *Journal of Experimental Child Psychology*, *162*, 209-224. https://doi.org/10.1016/j.jecp.2017.05.005

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, *17*(1), 99-103. https://doi.org/10.1037/0012-1649.17.1.99

Fodor, J. A. (1985). Precis of the modularity of mind. *Behavioral and Brain Sciences*, *8*(1), 1-5. https://doi.org/10.1017/S0140525X0001921X

Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2018). Reading your mind while you are reading—evidence for spontaneous visuospatial perspective-taking during a semantic categorization task. *Psychological Science*, *29*(4), 614-622. https://doi.org/10.1177/0956797617740973

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 459-473. https://doi.org/10.1098/rstb.2002.1218

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531-534. https://doi.org/10.1016/j.neuron.2006.05.001

Furlanetto, T., Becchio, C., Samson, D., & Apperly, I. (2016). Altercentric interference in level 1 visual perspective-taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(2), 158. https://doi.org/10.1037/xhp0000138

Gautam, S., Suddendorf, T., & Redshaw, J. (2021). When can young children reason about an exclusive disjunction? A follow up to. *Cognition*, *207*, 104507. https://doi.org/10.1016/j.cognition.2020.104507

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350-363. https://doi.org/10.1038/nrn3476

Gopnik, A., & Wellman, H.M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, *7*(1–2), 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. *Developmental Science*, *20*(5), e12445. https://doi.org/10.1111/desc.12445

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43-61. https://doi.org/10.1016/S0749-596X(03)00022-6

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*(5843), 1360-1366. https://doi.org/10.1126/science.1146282

Heyes, C. (2014a). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131-143. https://doi.org/10.1177/1745691613518076

Heyes, C. (2014b). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647-659. https://doi.org/10.1111/desc.12148

Hegedüs, A. M., & Király, I. (2022). Spontaneous attribution of underspecified belief of social partners facilitates processing shared information. *Scientific Reports*, *12*(1), 1-9. https://doi.org/10.1038/s41598-022-19569-8

Holland, C., Shin, S. M, & Phillips, J. (2021). Do you see what I see? A meta-analysis of the Dot Perspective Task. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43. Retrieved from https://escholarship.org/uc/item/7cs5r2xq

Hyde, D. C., Aparicio Betancourt, M., & Simon, C. E. (2015). Human temporal-parietal junction spontaneously tracks others' beliefs: A functional near-infrared spectroscopy study. *Human Brain Mapping*, *36*(12), 4831-4846. https://doi.org/10.1002/hbm.22953

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, *193*, 103950. https://doi.org/10.1016/j.cognition.2019.04.019

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133-156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct replication attempt of Southgate, Senju and Csibra (2007). *Royal Society Open Science*, *8*(8), 210190. https://doi.org/10.1098/rsos.210190

Kampis, D., & Kovács, Á. M. (2022). Seeing the World From Others' Perspective: 14-Month-Olds Show Altercentric Modulation Effects by Others' Beliefs. *Open Mind*, *5*, 189-207. https://doi.org/10.1162/opmi_a_00050

Kampis, D., & Southgate, V. (2020). Altercentric cognition: how others influence our cognitive processing. *Trends in Cognitive Sciences*, *24*(11), 945-959. https://doi.org/10.1016/j.tics.2020.09.003

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). The goggles experiment: Can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour*, *105*, 211-221. https://doi.org/10.1016/j.anbehav.2015.04.028

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32-38. https://doi.org/10.1111/1467-9280.00211

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25-41. https://doi.org/10.1016/s0010-0277(03)00064-7

Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false beliefs in 3-year-old children. *Proceedings of the National Academy of Sciences*, *115*(45), 11477-11482. https://doi.org/10.1073/pnas.1803505115

Knudsen, B., & Liszkowski, U. (2012a). Eighteen-and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science*, *15*(1), 113-122. https://doi.org/10.1111/j.1467-7687.2011.01098.x

Knudsen, B., & Liszkowski, U. (2012b). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, *17*(6), 672-691. https://doi.org/10.1111/j.1532-7078.2011.00105.x

Kovács, Á. M. (2016). Belief files in theory of mind reasoning. *Review of Philosophy and Psychology*, *7*(2), 509-527. https://doi.org/10.1007/s13164-015-0236-5

Kovács, Á. M., 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*, *9*(9), e106558. https://doi.org/10.1371/journal.pone.0106558

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830-1834. https://doi.org/10.1126/science.1190792

Kovács, Á. M., Téglás, E., & Csibra, G. (2021). Can infants adopt underspecified contents into attributed beliefs? Representational prerequisites of theory of mind. *Cognition*, *213*, 104640. https://doi.org/10.1016/j.cognition.2021.104640

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), e1503. https://doi.org/10.1002/wcs.1503

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110-114. https://doi.org/10.1126/science.aaf8110

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888-900. https://doi.org/10.1177/0956797617747090

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*, *46*, 97-111. https://doi.org/10.1016/j.cogdev.2017.09.001

Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1469-1482. https://doi.org/10.1037/0278-7393.21.6.1469

Lea, R. B., O'Brien, D. P., Fisch, S. M., Noveck, I. A., & Braine, M. D. (1990). Predicting propositional logic inferences in text comprehension. *Journal of Memory and Language*, *29*(3), 361-387. https://doi.org/10.1016/0749-596X(90)90005-K

Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, *24*(1), 65-78. https://doi.org/10.1016/j.tics.2019.11.004

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, *8*(12), 528-533. https://doi.org/10.1016/j.tics.2004.10.001

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*(3), 551-556. https://doi.org/10.1016/j.jesp.2009.12.019

Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives*, *10*(3), 184-189. https://doi.org/10.1111/cdep.12183

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*(3), 305-311. https://doi.org/10.1177/0956797612451469

Marshall, J., Gollwitzer, A., & Santos, L. R. (2018). Does altercentric interference rely on mentalizing?: Results from two level-1 perspective-taking tasks. *PloS ONE*, *13*(3), e0194101.https://doi.org/10.1371/journal.pone.0194101

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind?. *Trends in Cognitive Sciences*, *20*(5), 375-382. https://doi.org/10.1016/j.tics.2016.03.005

McKinnon, M. C., & Moscovitch, M. (2007). Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. *Cognition*, *102*(2), 179-218. https://doi.org/10.1016/j.cognition.2005.12.011

Meert, G., Wang, J., & Samson, D. (2017). Efficient belief tracking in adults: The role of task instruction, low-level associative processes and dispositional social functioning. *Cognition*, *168*, 91-98. https://doi.org/10.1016/j.cognition.2017.06.012

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, *22*(4), 280-293. https://doi.org/10.1016/j.tics.2018.02.001

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40-48. . https://doi.org/10.1016/j.cognition.2016.05.012

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*(2), 661-673. https://doi.org/10.1111/j.1467-8624.2010.01571.x

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297. https://doi.org/10.1037/0033-2909.132.2.297

Naughtin, C. K., Horne, K., Schneider, D., Venini, D., York, A., & Dux, P. E. (2017). Do implicit and explicit belief processing share neural substrates? *Human Brain Mapping*, *38*(9), 4760-4772. https://doi.org/10.1002/hbm.23700

Nijhof, A. D., Bardi, L., Brass, M., & Wiersema, J. R. (2018). Brain activity for spontaneous and explicit mentalizing in adults with autism spectrum disorder: An fMRI study. *NeuroImage: Clinical*, *18*, 475-484. https://doi.org/10.1016/j.nicl.2018.02.016

Nijhof, A. D., Brass, M., Bardi, L., & Wiersema, J. R. (2016). Measuring mentalizing ability: A within-subject comparison between an explicit and implicit version of a ball detection task. *PloS ONE*, *11*(10), e0164373. https://doi.org/10.1371/journal.pone.0164373

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255-258. https://doi.org/10.1126/science.1107621

O'Reilly, J. X. (2013). Making predictions in a changing world—inference, uncertainty, and learning. *Frontiers in Neuroscience*, *7*, 105. https://doi.org/10.3389/fnins.2013.00105

Perner, J. (1988). Developing semantics for theories of mind: From propositional attitudes to mental representation. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing theories of mind* (pp. 141–172). Cambridge University Press.

Perner, J., Huemer, M., & Leahy, B. (2015). Mental files and belief: A cognitive theory of how children represent belief and its intensionality. *Cognition, 145,* 77–88. https://doi.org/10.1016/j.cognition.2015.08.006

Perner, J., & Leahy, B. (2016). Mental files in development: Dual naming, false belief, identity and intensionality. *Review of Philosophy and Psychology, 7*(2), 491-508. https://doi.org/10.1007/s13164-015-0235-6

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *26*(9), 1353-1367. https://doi.org/10.1177/0956797614558717

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515-526. https://doi.org/10.1017/S0140525X00076512

Qureshi, A. W., Apperly, I. A., & 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*(2), 230-236. https://doi.org/10.1016/j.cognition.2010.08.003

Qureshi, A. W., & Monk, R. L. (2018). Executive function underlies both perspective selection and calculation in Level-1 visual perspective-taking. *Psychonomic Bulletin & Review*, *25*(4), 1526-1534. https://doi.org/10.3758/s13423-018-1496-8

Ragni, M., & Johnson-Laird, P. N. (2020). Reasoning about epistemic possibilities. *Acta Psychologica*, *208*, 103081. https://doi.org/10.1016/j.actpsy.2020.103081

Rakoczy, H. (2012). Do infants have a theory of mind? *British Journal of Developmental Psychology*, *30*(1), 59-74. https://doi.org/10.1111/j.2044-835X.2011.02061.x

Rakoczy, H. (2017). In defense of a developmental dogma: Children acquire propositional attitude folk psychology around age 4. *Synthese*, *194*(3), 689-707. https://doi.org/10.1007/s11229-015-0860-8

Redshaw, J. (2014). Does metarepresentation make human mental time travel unique?. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(5), 519-531. https://doi.org/10.1002/wcs.1308

Redshaw, J., & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, *26*(13), 1758-1762.https://doi.org/10.1016/j.cub.2016.04.062

Reverberi, C., Cherubini, P., Rapisarda, A., Rigamonti, E., Caltagirone, C., Frackowiak, R. S., ... & Paulesu, E. (2007). Neural basis of generation of conclusions in elementary deduction. *Neuroimage*, *38*(4), 752-762. https://doi.org/10.1016/j.neuroimage.2007.07.060

Reverberi, C., Pischedda, D., Burigo, M., & Cherubini, P. (2012). Deduction without awareness. *Acta Psychologica*, *139*(1), 244-253. https://doi.org/10.1016/j.actpsy.2011.09.011

Rubio-Fernández, P. (2017). The director task: A test of Theory-of-Mind use or selective attention?. *Psychonomic Bulletin & Review*, *24*(4), 1121-1128. https://doi.org/10.3758/s13423-016-1190-7

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science*, *24*(1), 27-33. https://doi.org/10.1177/0956797612447819

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

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology*, *80*(3), 201-224. https://doi.org/10.1006/jecp.2001.2633

Ruffman, T., & Perner, J. (2005). Do infants really understand false belief? Response to Leslie. *Trends in Cognitive Sciences*, *9*(10), 462-463. https://doi.org/10.1016/j.tics.2005.08.001

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255-1266. https://doi.org/10.1037/a0018729

Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., & Heyes, C. (2014). Avatars and arrows: Implicit mentalizing or domain-general processing?. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 929-937. https://doi.org/10.1037/a0035175

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16(2),* 235-239. https://doi.org/ 10.1016/j.conb.2006.03.001

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433. https://doi.org/10.1037/a0025458

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, *23*(*8),* 842-847. https://doi.org/10.1177/0956797612439070

Schneider, D., Nott, Z. E., & Dux, P. E. (2014). Task instructions and implicit theory of mind. *Cognition*, *133*(1), 43-47. https://doi.org/10.1016/j.cognition.2014.05.016

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition*, *162*, 27-31. https://doi.org/10.1016/j.cognition.2017.01.018

Schuwerk, T., Kampis, D., Baillargeon, R., Biro, S., Bohn, M., Byers-Heinlein, K., … Rakoczy, H. (2021, February 14). Action anticipation based on an agent's epistemic state in toddlers and adults. https://doi.org/10.31234/osf.io/x4jbm

Scott, R. M., & Baillargeon, R. (2014). How fresh a look? A reply to Heyes. *Developmental science*, *17*(5), 660-664. https://doi.org/10.1111/desc.12173

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237-249. https://doi.org/10.1016/j.tics.2017.01.012

Seow, T., & Fleming, S. M. (2019). Perceptual sensitivity is modulated by what others can see. *Attention, Perception, & Psychophysics*, *81*(6), 1979-1990. https://doi.org/10.3758/s13414-019-01724-5

Setoh, P., Scott, R.M., Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences*, *113*(47), 13360-13365. https://doi.org/10.1073/pnas.1609203113

Speiger, M. L., Langemeyer, L., Rakoczy, H., Liszkowski, U., & Grosse Wiesmann, C. (2021). *The Sandbox Task: A novel task to measure implicit and explicit Theory of Mind*. Poster presented at BCCCD 2021, Virtual.

Song, H. J., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology*, *44*(6), 1789-1795. https://doi.org/10.1037/a0013774

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

Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, *80*(4), 1172-1196. https://doi.org/10.1111/j.1467-8624.2009.01324.x.

Southgate, V. (2020). Are infants altercentric? The other and the self in early social cognition. *Psychological Review*, *127*(4), 505-523. https://doi.org/10.1037/rev0000182

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587-592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, *30*(3), 395-402. https://doi.org/10.1037/0012-1649.30.3.395

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580-586. https://doi.org/10.1111/j.1467-9280.2007.01943.x

Surtees, A. D., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology*, *30*(1), 75-86. https://doi.org/10.1111/j.2044-835X.2011.02063.x

Surtees, A., Apperly, I., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition*, *150*, 43-52. https://doi.org/10.1016/j.cognition.2016.01.014

Surtees, A., Samson, D., & Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97-105. https://doi.org/10.1016/j.cognition.2015.12.010

Symeonidou, I., Dumontheil, I., Ferguson, H. J., & Breheny, R. (2020). Adolescents are delayed at inferring complex social intentions in others, but not basic (false) beliefs: An eye-movement investigation. *Quarterly Journal of Experimental Psychology*, *73*(10), 1640-1659. https://doi.org/10.1177/1747021820920213

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632-1634. https://doi.org/10.1126/science.7777863

Tauzin, T., & Gergely, G. (2019). Variability of signal sequences in turn-taking exchanges induces agency attribution in 10.5-mo-olds. *Proceedings of the National Academy of Sciences*, *116*(31), 15441-15446. https://doi.org/10.1073/pnas.1816709116

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

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, *115*(34), 8491-8498. https://doi.org/10.1073/pnas.1804761115

Tomasello M., Carpenter M., Call J., Behne T., & Moll H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675-91. https://doi.org/10.1017/S0140525X05000129

Valle, A., Massaro, D., Castelli, I., & Marchetti, A. (2015). Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability. *Europe's Journal of psychology*, *11*(1), 112-124. https://doi.org/10.5964/ejop.v11i1.829

van Der Wel, R. P., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*(1), 128-133. https://doi.org/10.1016/j.cognition.2013.10.004

Van Overwalle, F., & Vandekerckhove, M. (2013). Implicit and explicit social mentalizing: dual processes driven by a shared neural network. *Frontiers in Human Neuroscience*, *7*, 560. https://doi.org/10.3389/fnhum.2013.00560

Wang, L. U., & Leslie, A. M. (2016). Is implicit theory of mind the 'Real Deal'? The own-belief/true-belief default in adults and young preschoolers. *Mind & Language*, *31*(2), 147-176. https://doi.org/10.1111/mila.12099

Ward, E., Ganis, G., & Bach, P. (2019). Spontaneous vicarious perception of the content of another's visual perspective. *Current Biology*, *29*(5), 874-880. https://doi.org/10.1016/j.cub.2019.01.046

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, *15*(6), 728-755. https://doi.org/10.1080/17405629.2018.1435413

Westra, E. (2017). Spontaneous mindreading: A problem for the two-systems account. *Synthese*, *194*(11), 4559-4581. https://doi.org/10.1007/s11229-016-1159-0

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128. https://doi.org/10.1016/0010-0277(83)90004-5

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

Supplementary Materials for Chapter 2

Additional analyses

Anticipatory looking task

S2.1. Familiarization phase: proportion of looking in the first anticipatory period

Experiment 1

Analyses of the familiarization phase revealed that the proportion of looking towards the target box was significantly higher than chance (0.5), for both the ambiguous and the unambiguous trials (ambiguous: Z=-3.56, p <.001, r=.650; unambiguous: Z=-3.87, p <.001, r=.701), indicating that, by the time the test phase started, participants developed reliable expectations regarding where the upcoming events will take place (see **Table S2.1**)**.** The proportion of looking was significantly higher for the later miscategorized than for the later properly categorized ambiguous trials (anticipatory period 1: Z=-2.09, p =.037, r=.381).

Table S2.1 *The mean proportion of looking towards the target box (correct anticipation) in the familiarization phase of the two experiments on the ambiguous and unambiguous colour trials, and separately for the two ambiguous trials, in the first anticipatory period (SD).*

	UNAMBIGUOUS	AMBIGUOUS	MISCATAMB	PROPCATAMB
Experiment 1	0.63 (0.16)	0.71 (0.20)	0.66 (0.20)	0.58 (0.23)
Experiment 2	0.74 (0.24)	0.72 (0.22)	0.77 (0.24)	0.71 (0.29)

Note: MISCATAMB denotes the ambiguous colour that was later miscategorized by the partner (in the test phase), PROPCATAMB is the other ambiguous colour. *N*s vary due to missing data (no valid trials). For the MISCATAMB trials N=27 in Experiment 1 and N=33 in Experiment 2. For the PROPCATAMB trials N=29 in Experiment 1 and N=32 in Experiment 2.

Experiment 2

Wilcoxon Signed Rank Tests revealed that the proportion of looking towards the target box was significantly higher than chance (0.5) in the familiarization phase, on both the ambiguous and unambiguous trials (ambiguous: Z=-4.08, p <.001, r=.700; unambiguous: Z=-4.22, p <.001, r=.724). There was no significant difference between the two ambiguous conditions (Z=-.77, p =.439, r=.132).

Comparison of Experiment 1 and Experiment 2

Comparison of the two experiments with respect to the proportion of looking in the first anticipatory period revealed a significant difference between the two: participants looked significantly more towards the target location (i.e. less towards the incorrect location) in Experiment 2 than Experiment 1, on both types of ambiguous trials prior the partner's picture selection (later miscategorized ambiguous trials: M_{Exp1}=0.66 vs M_{Exp2}=0.77, Z=-2.12, p= .034, r=0.265; later properly categorized ambiguous trials: M_{Exp1}=0.58 vs M_{Exp2}=0.71, Z=-2.13, p= .033, r=0.266), suggesting that it may have been easier to develop expectations regarding the partner's actions in the second than in the third experiment.

S2.2. Test phase: proportion of looking in the first anticipatory period

Proportion of looking results for the first anticipatory period of the test phase of are presented on **Figure S2.2** (left: Experiment 1; right: Experiment 2).

Experiment 1

Analysis of the proportion of looking data in the first anticipatory period revealed no significant main effect of condition (Wald χ^2=1.85, df=1, p= .174). There was, however, a tendency level main effect of block (Wald χ^2=6.43, df=3, p= .092) as well as a tendency level condition x block interaction (Wald χ^2=7.63, df=3, p= .054). As can be seen from the figure, the values were similar in the two conditions in the first three blocks. However, by the fourth block, participants started to look more towards the target box on the miscategorized than towards the incorrect box on the properly categorized ambiguous trials (z=2.57, M_{diff}= 0.18, Wald 95% CI [0.04-0.31], p_{adj}=.040).

Experiment 2

Analysis of the proportion of looking data in the first anticipatory period revealed no significant main effect of block (Wald χ^2=2.94, df=3, p= .401). There was, however, a significant main effect of condition (Wald χ^2=11.41, df=1, p<.001), and a tendency level condition x block interaction (Wald χ^2=7.23, df=3, p= .065). While the proportion of looking towards the incorrect box on the properly categorized ambiguous trials did not change substantially over time, the proportion of looking towards the target

box on the miscategorized trials increased markedly after the first block. Pairwise comparisons indicated significant differences between the two conditions from the second block on (miscategorized>properly categorized: block2 – z=3.13, M_{diff}= 0.25, Wald 95% CI [0.09-0.41], p_{adj}=.008; block3 – z=2.59, M_{diff}= 0.23, Wald 95% CI [0.05-0.40], p_{adj}=.040; block4 – z=3.04, M_{diff}= 0.23, Wald 95% CI [0.08-0.39], p_{adj}=.008).

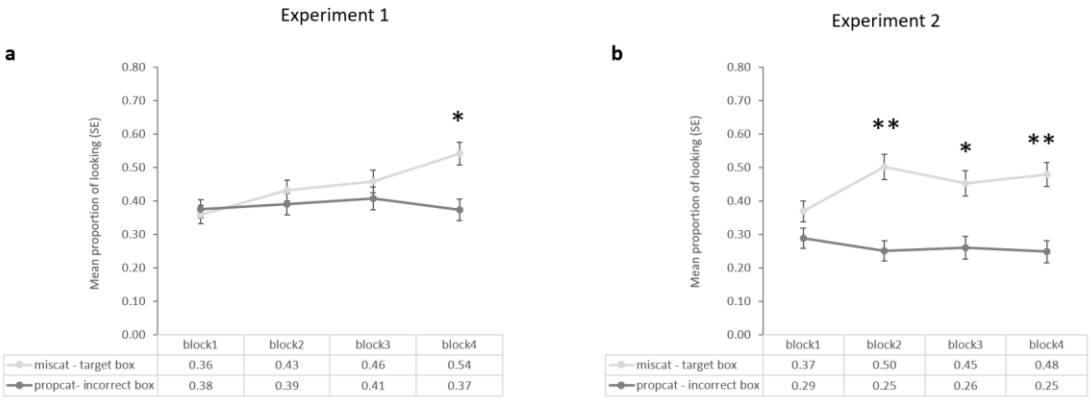


Figure S2.2. Changes in the proportion of looking towards the target box in the miscategorized condition (miscat: light grey line) versus the incorrect box in the properly categorized ambiguous condition (propcat: dark grey line) (a) in Experiment 1 (colour labels) and (b) Experiment 2 (colour matching) prior the partner's picture selection (anticipatory period 1), during the test phase of the anticipatory looking task. The figure displays the raw data. Analyses were run on the estimated marginal means. Error bars represent SE.
+: p<0.1; *: p<0.05, **: p<0.01

Comparison of Experiment1 and Experiment 2

Comparison of the two experiments along the proportion of looking data in the first anticipatory period yielded a significant main effect of condition (Wald χ^2=12.21, df=1, p<.001), block (Wald χ^2=8.28, df=3, p= .041) and a significant condition x block interaction (Wald χ^2=11.23, df=3, p= .011). In addition, there was also a significant main effect of experiment (Wald χ^2=4.31, df=1, p= .038) and a tendency level experiment x condition interaction (Wald χ^2=3.28, df=1, p= .070), resulting from the fact that participants in Experiment 1 looked more towards the incorrect box on the properly categorized ambiguous trials than participants in Experiment 2. Neither the experiment x block (Wald χ^2=1.99, df=3, p=.574) nor the experiment x condition x block interaction was significant (Wald χ^2=2.95, df=3, p=.400), indicating that, the way in which the proportion of looking evolved in the two conditions over

time, did not differ between the two groups. Importantly, the difference with respect to the proportion of looking on the properly categorized ambiguous trials was already present by the time the test phase started suggesting that the observed effect might not have been related to the manipulations of the test phase but rather pre-existing differences in the strength of expectations (see the comparison of the familiarization phase above).

Performing the same analyses for the two 'update' subgroups yielded a significant main effect of condition (Wald χ^2=14.10., df=1, p<.001), block (Wald χ^2=17.10., df=3, p= .001) and experiment (Wald χ^2=5.853, df=1, p= .016) as well as a significant condition x block interaction (Wald χ^2=12.95, df=3, p= .005) in the first anticipatory period, but no significant experiment x condition (Wald χ^2=1.47, df=3, p=.306), experiment x block (Wald χ^2=5.27, df=3, p=.153) or experiment x condition x block interaction (Wald χ^2=5.39, df=3, p=.145), reflecting the fact that members of the 'update' group in Experiment 1 tended to look more towards both the target box on the miscategorized and the incorrect box on the properly categorized ambiguous trials than members of the 'update' group in Experiment 2.

Altogether findings from the first anticipatory period indicate that not only did participants revise their expectations regarding the other's behaviour but soon after they did so they started to use what they had learnt about their partner to predict her actions from the earliest possible timepoint.

S2.3. Proportion of looking in the first anticipatory period of the test phase: subgroup analyses

Figure S2.3 presents the proportion of looking results for the first anticipatory period of the test phase in the 'update' (left panel) and 'noupdate' (right panel) subgroups of Experiment 1 and Experiment 2.

Experiment 1

Analysis of the proportion of looking data in the 'update' group yielded a significant main effect of condition (Wald χ^2=3.90, df=1, p= .048) and block (Wald χ^2=42.60, df=3, p= .061). There was also a significant condition x block interaction (Wald χ^2=15.65, df=3, p= .001). While the proportion of looking towards the incorrect box in the properly categorized ambiguous condition did not change substantially with time, the proportion of looking towards the target box in the miscategorized condition increased sharply over the course of trials. Pairwise comparisons revealed that the difference between the two conditions became significant by the fourth block (miscategorized> properly categorized: z=3.43, M_{diff}= 0.37, Wald 95% CI [0.16-0.58], p_{adj}=.004).

For the 'noupdate' group, neither the main effect of condition (Wald χ^2=0.060, df=1, p= .807) or block (Wald χ^2=1.49, df=3, p= .686) nor the condition x block interaction was significant (Wald χ^2=1.41, df=3, p= .704). The proportion of looking was relatively high (~0.30-0.40) in both conditions, from the beginning of the test phase, throughout the four blocks.

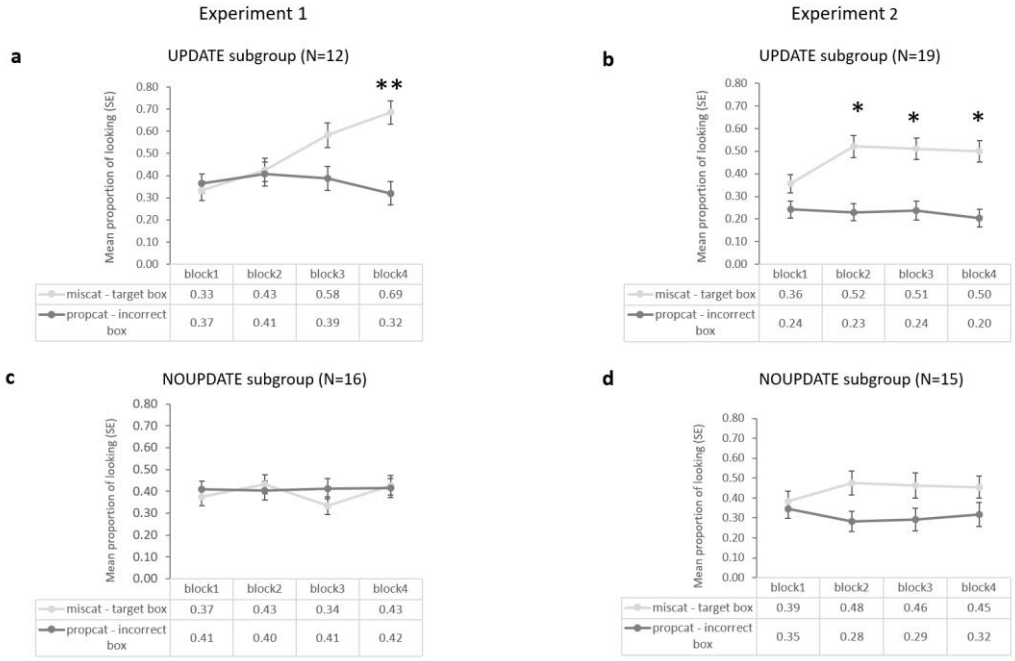


Figure S2.3. Changes in the proportion of looking towards the target box in the miscategorized condition (miscat: light grey line) versus the incorrect box in the properly categorized ambiguous condition (propcat: dark grey line) in the UPDATE (a, c) and NOUPDATE (b, d) subgroups of Experiment 1 (colour labels, left) and Experiment 2 (colour matching, right), prior the partner's picture selection. The subgroups were created on the basis of participants' performance on the other-perspective trials of explicit perspective-taking task. The figure displays the raw data. Error bars represent SE.
+: p<0.1; *: p<0.05, **: p<0.01

Experiment 2

Analysis of the proportion of looking data in the 'update' group revealed no significant main effect of block (Wald χ^2=4.86, df=3, p= .183). There was, however, a significant main effect of condition (Wald χ^2=10.82, df=1, p= .001), showing that participants generally looked more towards the target box on the miscategorized than towards the incorrect box on the properly categorized ambiguous trials, throughout the test phase of the task. Although the difference between the two conditions became pronounced only after the first block, with pairwise comparisons indicating a significant difference between the two conditions only in the subsequent blocks (miscategorized>properly categorized: block2 – z=2.64, M_{diff}= 0.29, Wald 95% CI [0.08-0.50], p_{adj}=.032; block3 – z=2.63, M_{diff}= 0.27, Wald 95%

CI [0.07-0.46], p_{adj}=.036; block4 – z=2.92, M_{diff}= 0.29, Wald 95% CI [0.09-0.8], p_{adj}=.016), the condition x block interaction was not significant (Wald χ^2=5.58, df=3, p= .134).

With respect to the 'noupdate' group, despite participants tended to look more towards the target box on the miscategorized than towards the incorrect box on the properly categorized ambiguous trials, none of the effects were significant (condition: Wald χ^2=2.24, df=1, p= .134; block: Wald χ^2=0.21, df=3, p= .976; condition x block: Wald χ^2=2.44, df=3, p= .486).

Taken together, these results show that, just like in case of the second anticipatory looking period, group-level findings were driven by the looking behaviour of the 'update' subgroup.

S2.4. First look ratios in the test phase

First look ratios of the two conditions in the four blocks of the test phase are presented on **Figure S2.4** (upper panel: anticipatory period 1, lower panel: anticipatory period 2).

Experiment 1

With respect to the first anticipatory period, the analysis revealed a tendency level effect of block (Wald χ^2=7.45, df=3, p= .059), but no significant main effect of condition (Wald χ^2=0.01, df=1, p= .938) or condition x block interaction (Wald χ^2=4.46, df=3, p= .216), reflecting a modest increase in the number of first looks towards the rule-incongruent locations (i.e. target box on the miscategorized and incorrect box on the properly categorized ambiguous trials) in the two ambiguous colour conditions over the course of trials. For the second anticipatory period, results were similar to the ones obtained in the analysis of the proportion of looking data: there was a significant main effect of condition (Wald χ^2=10.06, df=1, p= .002) and block (Wald χ^2=9.90, df=3, p= .019), as well as a significant condition x block interaction (Wald χ^2=20.35, df=3, p< .001), resulting from a sharp increase in the number of first looks towards the target box on the miscategorized and a slight decrease in the number of first looks towards the incorrect box on the properly categorized ambiguous trials, following the first block. Pairwise comparisons indicated a significant difference between the two conditions from the second block on (miscategorized>properly categorized: block2 – z=2.90, M_{diff}= 0.20, Wald 95% CI [0.06-0.33], p_{adj}=.020; block3 – z=2.76, M_{diff}= 0.21, Wald 95% CI [0.06-0.36], p_{adj}=.020; block4 – z=3.94, M_{diff}= 0.28, Wald 95% CI [0.14-0.42], p_{adj}<.001).

Experiment 2

Analysis of the first looks in the first anticipatory period yielded a significant main effect of condition (Wald χ^2=8.31, df=1, p= .004) but no significant main effect of block (Wald χ^2=0.430, df=3, p= .934) or condition x block interaction (Wald χ^2=1.81, df=3, p= .612), showing that participants tended to direct their first look towards the target box on the miscategorized trials more often than their first look towards the incorrect box on the properly categorized ambiguous trials, from the beginning of the task, with no substantial change in the number of first looks during the test phase trials. With respect to the second anticipatory period, again, there was a significant main effect of condition (Wald χ^2=20.21, df=1, p< .001) but no significant main effect of block (Wald χ^2=4.65, df=3, p= .200). There was, however, a significant condition x block interaction (Wald χ^2=26.14, df=3, p< .001), resulting from a sharp and steady increase in the number of first looks towards the target box on the miscategorized trials and a drop in the number of first looks towards the incorrect box on the properly categorized ambiguous trials, following the first block. Just like in Experiment1, pairwise comparisons indicated a significant difference between the two conditions from the second block on (miscategorized>properly categorized: block2 – z=4.86, M_{diff}= 0.34, Wald 95% CI [0.20-0.48], p_{adj}<.001; block3 – z=5.22, M_{diff}= 0.36, Wald 95% CI [0.22-0.50], p_{adj}<.001; block4 – z=5.00, M_{diff}= 0.43, Wald 95% CI [0.26-0.60], p_{adj}<.001). Importantly, in this experiment, the ratio of first looks exceeded 0.50 by the end of the test phase, although the difference was not statistically significant (block4: Z=1.36, p=0.174, r=0.23).

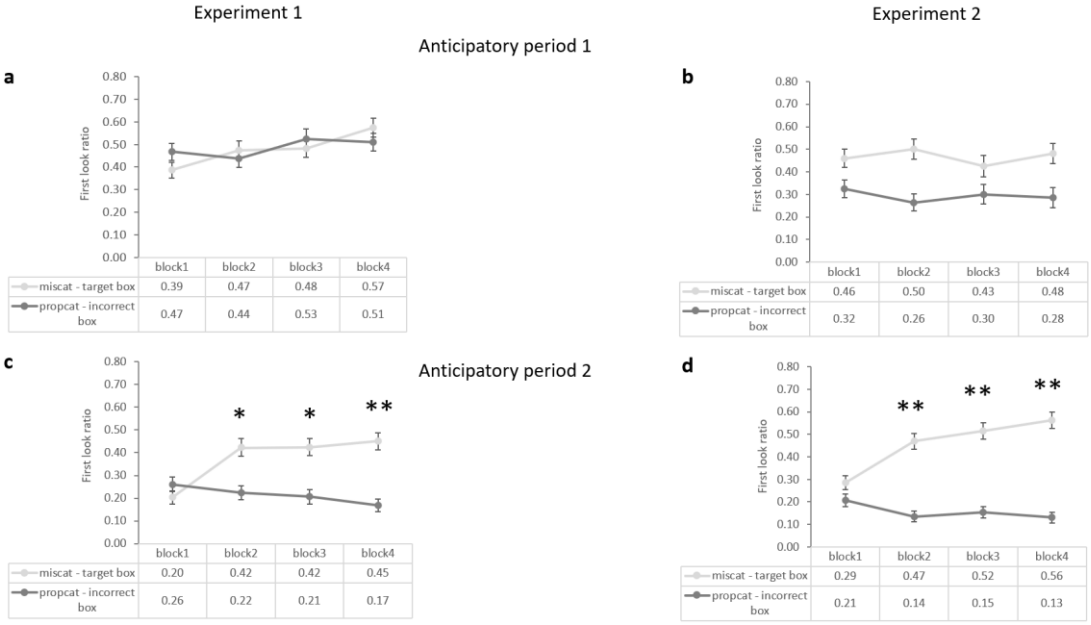


Figure S2.4. Changes in the frequency of first looks directed towards the target box in the miscategorized condition (miscat: light grey line) versus the incorrect box in the properly categorized ambiguous condition (propcat: dark grey line) (a, b) prior the partner's picture selection (anticipatory period 1), and (c, d) box selection (anticipatory period 2), during the test phase of Experiment1 (colour labels, left) and Experiment 2 (colour matching, right). The figure displays the raw data. Analyses were run on the estimated marginal means. Error bars represent SE. +: p<0.1; *: p<0.05, **: p<0.01

Comparison of Experiment1 and Experiment 2

For the first anticipatory period, comparison of the two experiments revealed a significant main effect of condition (Wald χ^2=5.46, df=1, *p*=.019), experiment (Wald χ^2=12.93, df=1, *p*< .001) as well as a significant experiment x condition interaction (Wald χ^2=5.80, df=3, *p* =.016), resulting from the fact that participants in Experiment 2 looked much less often first towards the incorrect box on the properly categorized ambiguous trials than participants in Experiment 1, throughout the test phase of the task. Neither the main effect of block (Wald χ^2=4.98, df=3, *p*= .173) nor any of the other interactions were significant (condition x block: Wald χ^2=5.58, df=3, *p*= .134; experiment x block - Wald χ^2=4.57, df=3, *p*=.206; experiment x condition x block interaction: Wald χ^2=0.35, df=3, *p*=.951), reflecting the lack of change in the level of first look ratios over the course of trials and generally similar pattern in Experiment 1 and Experiment 2.

With respect to the second anticipatory period, the analysis yielded a significant main effect of condition (Wald χ^2=30.27, df=1, *p*<.001), block (Wald χ^2=11.32, df=3, *p*= .010) and a significant condition x block interaction (Wald χ^2=39.84, df=3, *p*< .001), but no significant main effect of

experiment (Wald χ^2=0.16, df=1, *p*= .694). There was, however, a tendency level experiment x condition interaction (Wald χ^2=3.63, df=3, *p* =.057): participants in Experiment 2 tended to direct their first look towards the target box on the miscategorized trials more often than participants in Experiment 1, in general, from the beginning of the task. None of the other interactions were significant, reflecting that, in the two experiments the first looks evolved in a similar way in the two conditions during the test phase of the task (experiment x block - Wald χ^2=1.36, df=3, *p*=.716; experiment x condition x block interaction - Wald χ^2=0.46, df=3, *p*=.997).

Altogether, results of the first look analyses corroborate the findings with our proportion of looking measure: that participants spontaneously updated the other's mental state and started to revise their expectations regarding her future behaviour only after the observation of a few actions that were incongruent with their original assumptions about how she perceives/categorizes the colours.

S2.5. Test phase: proportion of looking towards the incorrect boxes on the unambiguous trials in the second anticipatory period

To investigate whether the observed changes in the looking behaviour indeed reflect the revision of the original assumptions regarding how the partner perceives the specific colour miscategorized by the partner or simply an increase in the tendency to look more towards the specific target box, for example, due to the change in the statistical regularities (which box lights up more often), we also ran the main analyses for the two unambiguous colours, comparing the proportion of looking and the first looks towards the incorrect box on the „unambiguous version of the miscategorized colour (blue for blueish, green for greenish) and the 'other unambiguous colour' trials.

Experiment 1

Figure S2.5a and **Figure S2.5c** depicts the mean proportion of looking and the first looks towards the incorrect box on the unambiguous trials in the four blocks of the test phase, prior to the partner's box selection, respectively. Analysis of the proportion of looking data revealed no significant main effect of condition (Wald χ^2=2.46, df=1, *p*= .117) or block (Wald χ^2=1.03, df=3, *p*= .794). There was however a significant condition x block interaction (Wald χ^2=13.97, df=3, *p*= .003), resulting from a slight increase in the proportion of looking towards the incorrect box on the unambiguous version of the

miscategorized and a modest decrease in the proportion of looking towards the incorrect box on the other unambiguous colour trials, after the first block. Importantly, proportion of looking towards the incorrect box started to decrease after this initial increase on the 'unambiguous miscategorized' trials. Pairwise comparisons indicated a significant difference between the two conditions only in the second (z=3.48, M_{diff}= 0.10, Wald 95% CI [0.04-0.16], p_{adj}<.001), but not in the subsequent blocks (block3 – z=1.94, M_{diff}= 0.07, Wald 95% CI [0.001-0.13], p_{adj}=.212; block4 – z=0.34, M_{diff}= 0.01, Wald 95% CI [-0.07-0.10], p_{adj}=1.00).

The pattern was similar for the first looks. Analysis indicated no significant main effect of condition (Wald χ^2=1.83, df=1, p= .177) and block (Wald χ^2=2.87, df=3, p= .412) but there was a significant condition x block interaction (Wald χ^2=20.32, df=3, p< .001). Pairwise comparisons, again, indicated a significant difference between the two conditions only in the second block, after adjusting for multiple comparisons (z=2.92, M_{diff}= 0.10, Wald 95% CI [0.04-0.20], p_{adj}<.001).

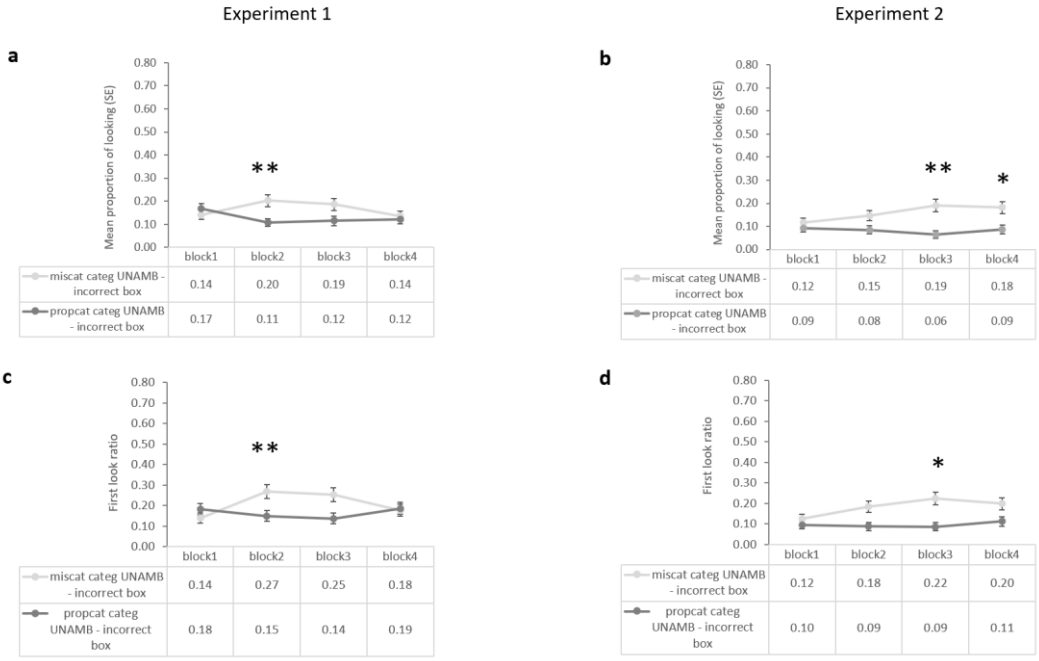


Figure S2.5. Changes in the proportion of looking (upper panel) and in the frequency of first looks (lower panel) directed towards the incorrect box on trials where the frame had the unambiguous version of the miscategorized colour (miscat categ UNAMB: light grey line) versus the other unambiguous colour trials (propcat categ UNAMB: dark grey line) (a, c) in Experiment 1 (colour labels) and (b, d) Experiment 2 (colour matching) prior the partner's box selection (anticipatory period 2), during the test phase of the anticipatory looking task, in the total sample. The figure displays the raw data. Analyses were run on the estimated marginal means. Error bars represent SE. +: p<0.1; *: p<0.05, **: p<0.01

Analyses revealed a similar pattern in the 'update subgroup' for both measures (proportion of looking – condition: Wald χ^2=0.55, df=1, p= .458; block: Wald χ^2=4.32, df=3, p= .229; condition x block: Wald χ^2=32.24, df=3, p< .001; first look – condition: Wald χ^2=0.81, df=1, p= .369; block: Wald χ^2=0.006, df=3, p= 1.00; condition x block: Wald χ^2=19.14, df=3, p< .001): a modest increase after the first block,

followed by a decrease in the proportion of looking and number of first looks towards the incorrect box on trials where the frame's colour was the unambiguous version of the miscategorized colour with the proportions/first look ratios remaining rather low in all four blocks (proportion of looking: <0.1; first look<0.20). The difference between the two unambiguous conditions was significant only in the second block and only in one of the measures, after adjusting for multiple comparisons (proportion of looking – block2: z=3.29, M_{diff}= 0.08, Wald 95% CI [0.03-0.13], p_{adj}<.001; first look – block2: z=1.92, M_{diff}= 0.10, Wald 95% CI [0.00-0.21], p_{adj}=.188). With respect to the 'noupdate' subgroup, despite the pattern was similar to the one observed in the update subgroup (and the total sample), the none of the effects was significant in the proportion of looking analyses (condition: Wald χ^2=1.90, df=1, p= .198; block: Wald χ^2=1.42, df=3, p= .700; condition x block: Wald χ^2=1.53, df=3, p= .201). Analysis of the first look data revealed a significant condition x block interaction (Wald χ^2=12.56, df=3, p= .006), but no main effect of condition (Wald χ^2=0.6, df=1, p= .427) or block (Wald χ^2=4.22, df=3, p= .238), with pairwise comparisons indicating a tendency level difference between the two unambiguous colour conditions only in the second block, after adjusting for multiple comparisons (z=2.33, M_{diff}= 0.14, Wald 95% CI [0.03-0.26], p_{adj}=.060). Importantly the mean proportion of looking and first look ratio was generally higher in this than in the update subgroup on trials where the frame had the unambiguous version of the miscategorized colour (proportion of looking: <0.28; first look<0.33).

Taken together, these results make it rather unlikely that the observed change in participants' looking behaviour on the miscategorized trials, reflected simply a tendency to look more towards the box that lit up more frequently, as a result of our manipulation.

Experiment 2

The proportion of looking and the first look results for the two unambiguous conditions in the second anticipatory period of the four test phase blocks are presented on **Figure S2.5b** and **Figure S2.5d**, respectively.

Analysis of the proportion of looking data yielded a significant main effect of condition (Wald χ^2=7.97, df=1, p= .005) but no significant main effect of block (Wald χ^2=1.63, df=3, p= .652). In addition, there was a tendency level condition x block interaction (Wald χ^2=6.47, df=3, p= .091), due to a slight increase in the proportion of looking towards the incorrect box on trials where the frame had the unambiguous version of the miscategorized colour trials (but not on the other unambiguous colour trials), starting after the first block. The difference between the two conditions became significant by the third block

(block3 – z=3.03, M_{diff}= 0.12, Wald 95% CI [0.04-0.20], p_{adj}=.008; block4 – z=2.65, M_{diff}= 0.10, Wald 95% CI [0.02-0.17], p_{adj}=.032).

With respect to the first look ratios, the analysis revealed a significant main effect of condition (Wald χ^2=9.50, df=1, p= .002) but no significant main effect or block (Wald χ^2=2.60, df=3, p= .458). Despite a modest increase in the number of first looks towards the incorrect box on the unambiguous miscategorized colour trials (which was not present in the other unambiguous colour condition), the condition x block interaction was not significant either (Wald χ^2=3.39, df=3, p= .335). Pairwise comparisons, however, indicated a significant difference between the two conditions, in the third block (z=2.98, M_{diff}= 0.14, Wald 95% CI [0.04-0.23], p_{adj}=.016).

Regarding the 'update subgroup', both analyses yielded a significant main effect of condition (proportion of looking: Wald χ^2=4.09, df=1, p= .043; first look: Wald χ^2=4.38, df=1, p= .036): participants tended to look somewhat more and direct their first look slightly more frequently towards the incorrect box on trials where the frame had the unambiguous version of the miscategorized colour, with no substantial change over time (proportion of looking – block: Wald χ^2=1.17, df=3, p= .761; condition x block: Wald χ^2=1.74, df=, p= .628; first look - block: Wald χ^2=1.41, df=3, p= .703; condition x block: Wald χ^2=0.91, df=3, p= .824). Importantly, however, the proportions as well as the first look ratios remained rather low even on these trials, throughout the experiment (proportion of looking <0.10; first look<0.14). With respect to the 'noupdate subgroup', the pattern was somewhat different. Analyses yielded a significant main effect of condition (Wald χ^2=5.15, df=1, p= .023) but no significant main effect of block on the proportion of looking (Wald χ^2=4.8, df=3, p= .223) and a significant main effect of both the condition (Wald χ^2=5.84, df=1, p= .016) and the block on the first look data (Wald χ^2=8.80, df=3, p= .032). Importantly, the condition x block interaction was also significant or marginally significant, in both cases (proportion of looking: Wald χ^2=7.46, df=3, p= .059; first look: Wald χ^2=8.3, df=3, p= .040), resulting from the fact that participants tended to look more and look first more frequently towards the incorrect box on the unambiguous miscategorized colour (but not on the other unambiguous colour) trials after the first block. Pairwise comparisons revealed that the difference between the two conditions became significant by fourth block (proportion of looking - z=2.94, M_{diff}= 0.23, Wald 95% CI [0.08-0.38], p_{adj}=.012; first look – z=3.98, M_{diff}= 0.33, Wald 95% CI [0.17-0.49], p_{adj}=.080). Notably, as in Experiment 1, the mean proportion of looking and first look ratios were much higher on the unambiguous miscategorized colour trials in this than in the update subgroup (proportion of looking: <0.36; first look<0.41).

Altogether, the pattern of results suggests that the observed slight increase in the looking towards the 'unambiguous miscategorized' box resulted from the 'noupdate' participants' difficulties to find a satisfying explanation for the change in the partner's behaviour. Hence, it is unlikely that the change

in participants' anticipatory looking behaviour on the ambiguous trials reflected merely a reaction to the change in the statistical regularities (i.e. which specific box lit up more often).

S2.6. Verbal reports

In Experiment 1, 25 out of the 30 participants noticed the change in the 'partner's' behaviour (83.33%), and 16 out of the 25 were able to specify exactly how the other's behaviour deviated from the expected (64%). Altogether 18 of the 25 participants gave mentalistic accounts for the change, such as the other perceives the colours in a different way than they themselves or was shown different colours/pictures (72%). There were, eight participants out of the 30 who expressed serious doubts that they were actually playing with their partner (as indicated by <3 points given for the second question). These doubts, however, emerged during the experiment, as a result of how the other behaved during the task, and did not reflect pre-existing beliefs regarding the 'partner's' presence/absence, according to the participants' reports. Therefore, their data was not excluded from the analyses. In the 'update' group (n=12) 9 of the 12 participants (75%) could specify at the end which colour was miscategorized/misperceived by the other. Altogether 8 gave a mentalistic explanation for the change in her behaviour (67%) and 3 expressed serious doubts that they were playing with another human, as a result of how their 'partner' 'behaved' during the experiment (25%). In the 'noupdate' group (n=16) these numbers were 6 (37.5%), 8 (50%) and 5 (37.5%), respectively.

In Experiment 2 29 out of the 34 participants noticed the change in the 'partner's' behaviour (85%). Out of these 29, 22 was able to specify which of the two ambiguous colours was miscategorized/misperceived by the other (76%). Altogether 25 of the 29 participants gave mentalistic accounts for the observed behaviour (86%). Six out of the 34 participants expressed serious doubts that they were playing with their partner/another human. For reasons mentioned above, their data was not excluded from the analyses. In the 'update' group 17 out of the 19 participants could specify exactly how the other's behaviour deviated from the expected (89%), 16 gave mentalistic accounts for the observed change (84%). Of the 19 participants, there was only one who expressed serious doubts about playing with another human. In the 'noupdate' group (n=15) these numbers were 10 (66%), 9 (60%) and 5 (33%), respectively.

The two experiments did not differ in terms of how many participants could specify which of the two ambiguous colours was miscategorized/misperceived by the other ($\chi^2(1, N=64)=0.855$, $p=.355$, *Cramér's V*= .115, n=16 versus n=22) or how many gave mentalistic accounts for the unexpected

behaviour of the other at the end of the experiment ($\chi^2(1, N=64)=1.33$, $p=.250$, *Cramér's V*= .144, n=18 versus n=25).

S2.7. Anticipatory looking task: proportion of valid and noanticipation trials

Comparison of the two experiments

The proportion of valid trials and the proportion of trials on which participants did not anticipate towards either of the two boxes (labelled as 'noanticipation' trials) is presented in **Table S2.2**.

Table S2.2. *The proportion of valid and noanticipation trials in the anticipatory looking task Experiment 1 and Experiment 2*

	Experiment 1		Experiment 2	
	antper 1	antper 2	antper 1	antper 2
Proportion of valid trials – miscategorized (SD)	0.67 (0.29)	0.75 (0.30)	0.57 (0.27)	0.74 (0.25)
Proportion of valid trials – properly categorized (SD)	0.66 (0.29)	0.74 (0.31)	0.57 (0.25)	0.72 (0.29)
Proportion of noanticipation trials - miscategorized (SD)	0.32 (0.29)	0.25 (0.30)	0.42 (0.27)	0.25 (0.25)
Proportion of noanticipation trials - properly categorized (SD)	0.33 (0.29)	0.25 (0.31)	0.43 (0.25)	0.27 (0.29)
Low anticipators - miscategorized (N)	8	5	12	6
Low anticipators - properly categorized ambiguous condition (N)	7	5	10	7

Note: Valid trials are trials where the number of missing datapoints was > 50% and the participant anticipated towards either of the two boxes in the given anticipatory period. The proportion of 'noanticipaton' trials was calculated by dividing the number of trials where the participant did not anticipate towards either of the two boxes in the given time window/ 32. Low anticipators are participants who anticipated in < 50% of the trials.
antper1: anticipatory period 1 (before the partner's picture selection)
antper2: anticipatory period 2 (before the partner's box selection)

Analyses indicated that the two experiments did not differ significantly in the proportion of valid trials (anticipatory period 1 - miscategorized: Z=-1.44, *p*=.149, *r*=.180; properly categorized ambiguous: Z=-1.46, *p*=.144, *r*=.138; anticipatory period 2 - miscategorized: Z=-0.67, *p*=.504, *r*=.084; properly categorized ambiguous: Z=-0.63, *p*=.533, *r*=.079) or the proportion of 'noanticipation' trials in either of the two conditions and anticipatory periods (anticipatory period 1 - miscategorized: Z=-1.41, *p*=.157, *r*=.176; properly categorized ambiguous: Z=-1.52, *p*=.130, *r*=.190; anticipatory period 2 - miscategorized: Z=-0.49, *p*=.626, *r*=.061; properly categorized ambiguous: Z=-0.59, *p*=.555, *r*=. 074). Both in Experiment 1 there were N=5-5 participants who anticipated <50% in the second time window on both the miscategorized and the properly categorized ambiguous trials. In Experiment 2 these numbers were 6 and 7, respectively. These results indicate that the differences between the two

experiments were not due to general differences between the two groups in how motivated participants were to track their partner's actions.

In Experiment 2 there was a significant difference between the 'update' and the 'noupdate' group with respect to the proportion of 'noanticipation' trials in the second anticipatory period, i.e. on how many trials participants did not look towards any of the boxes prior the 'partner's' box selection (see **Table S2.3**). 'Noupdate' participants anticipated significantly less on both the miscategorized (Z=-2.84, p=.005, r=0.487) and the properly categorized ambiguous trials (Z=-2.88, p=.004, r=0.494) than members of the 'update' group. A similar trend could be observed for the first anticipatory period, the difference was, however, much smaller hence not significant for this time window (miscategorized: Z=-0.61, p=0.543, r= .105; properly categorized: Z=-0.99, p=0.322, r=.170). No such differences were present for Experiment 1, in either of the two time windows (all Zs < -0.73, all ps> .460). Altogether, the results suggest that, in Experiment2, differences between the two subgroups may have been - at least partially - the result of pre-existing differences in participants' propensity to predict/track the behaviour of the other.

Table S2.3. *The ratio of 'noanticipation' trials in the 'update' and 'noupdate' subgroups of Experiment 1 and Experiment 2 in the two anticipatory periods*

	Experiment 1		Experiment 2	
	'update' (N=12)	'noupdate' (N=16)	'update' (N=19)	'noupdate' (N=15)
Anticipatory period 1				
Proportion of noanticipation trials - miscategorized (SD)	0.34 (0.28)	0.31 (0.33)	0.39 (0.25)	0.44 (0.30)
Proportion of noanticipation trials - properly categorized (SD)	0.35 (0.26)	0.30 (0.31)	0.39 (0.24)	0.47.(0.26)
Anticipatory period 2				
Proportion of noanticipation trials - miscategorized (SD)	0.23 (0.30)	0.29 (0.32)	0.14 (0.17)	0.38 (0.28)
Proportion of noanticipation trials - properly categorized (SD)	0.22 (0.30)	0.29 (0.35)	0.3 (0.17)	0.45 (0.32)

Note: noanticipation trials are trials on which participants provided data but did not look at either of the two target locations

Supplementary Materials for Chapter 3

S3.1 RT data exclusions

Trials in which the wrong button or no button was pressed and reaction times were more than two standard deviations from the condition means of each participant (calculated separately for the two beliefs) were considered invalid and were excluded from the RT analyses.

In Experiment 1a this meant the exclusion of 5.53% of the RT data in the TB actual, 7.94% in the TB possible and 8.66% in the TB impossible trials and 7.21% of the RT data in the UB actual, 8.89% in the UB possible and 6.97% in the UB impossible trials. In Experiment 1b these numbers were 4.93%,10.60%, 8.90% and 5.50%,6.81%, 5.78%, respectively.

In Experiment 2 it meant the exclusion of 5.72 % of the RT data in the TB actual, 7.15% in the TB possible and 6.97% in the TB impossible trials and 6.43% of the RT data in the UB actual, 6.79% in the UB possible and 5.89% in the UB impossible trials.

In Experiment 3 7.12% of the RT data was excluded in the TB actual, 7.29% in the TB possible and 7.73% in the TB impossible trials. On the UB trials these numbers were 6.60%, 7.82% and 7.21%, respectively.

In Experiment 4 5% of the RT data was excluded in the TB possible and 3.39% in the TB impossible trials and 1.79% and 2.50% of the RT data in the UB possible and impossible trials of the 'ball absent' trials. In case of the 'ball present' trials applying our data exclusion criteria meant the exclusion of 3.21% of the RT data in the TB actual, 5.36% in the TB possible and 4.29% in the TB impossible trials and 2.50% of the RT data in the UB actual, 1.79% in the UB possible and 1.43% in the UB impossible trials.

S3.2 The effect of spatial position

S3.2.1 Experiment 1a

Reaction times

Results of a 2 x 2 x 3 repeated measures ANOVA, with position (central vs peripheral), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors, revealed that the observed effects did not depend on the actual spatial position occupied by the 'possible' box. In specific, the analysis yielded a significant main effect of belief ($F(1, 25)=9.13$, $p= .006$, $\eta_p^2= .268$), location type ($F(1.35, 33.68)=16.83$, $p< .001$, $\eta_p^2= .402$) as well as a significant belief x location type interaction ($F(2, 50)=4.63$, $p= .014$, $\eta_p^2= .156$), but, as can be seen on **Figure S3.1a** and **S3.1b**, although

participants were generally slower in detecting changes at the peripheral positions, compared to the central ones (position main effect: $F(1, 25)=99.99$, $p< .001$, $\eta_p^2= .800$), the pattern of RT results was the same for the two spatial locations, with neither the position x location type ($F(2, 50)=1.44$, $p= .247$, $\eta_p^2= .054$), nor position x belief ($F(1, 25)=0.17$, $p= 682$, $\eta_p^2= .007$) or the position x belief x location type interaction being significant ($F(2, 50)=0.88$, $p= .419$, $\eta_p^2= .034$).

Miss ratios

As can be seen on **Figure S3.1c** and **S3.1d**, the pattern of findings observed in the main analyses resulted mainly from participants' performance on the 'peripheral'-change trials, with Friedman tests indicating a marginally significant difference between the three location types for the TB ($\chi^2(2)=5.77$, $p= .056$, Kendall's $W=0.111$; follow-up Wilcoxon Signed Rank Tests: all Zs< 2.09, all p_{adj}s> .111) but not for the UB trials ($\chi^2(2)=0.298$, $p= .862$, Kendall's $W=0.006$), indicating that participants directed less attention to the actual location when the agent was uncertain regarding the location of the ball and this would have required more effort due to the box's spatial position. For changes at the central positions, a similar pattern could be observed for the two beliefs: no miss on the actual versus a few on the possible and the impossible trials, suggesting a modest actual bias when the box where the ball hid occupied a central position. The difference between the three types of locations was again marginally significant for the TB ($\chi^2(2)=5.82$, $p= .055$, Kendall's $W=0.112$; follow-up Wilcoxon Signed Rank Tests: all Zs< 2.04, all p_{adj}s> .123) but not for the UB trials ($\chi^2(2)=2.25$, $p= .325$, Kendall's $W=0.034$). Paralleling the findings in reaction times, miss ratios were generally higher (i.e. change detection performance was worse) for changes at the peripheral locations compared to the central ones, with the differences being significant, for all three location types both on the TB trials (actual: $Z=-2.53$, $p_{adj(3)}= .012$, $r=0.496$; possible: $Z=-2.88$, $p_{adj(3)}= .012$, $r=0.564$; impossible: $Z=-2.76$, $p_{adj(3)}= .012$, $r=0.541$) and for the actual ($Z=-2.46$, $p_{adj(3)}= .028$, $r=0.482$) and impossible location on the UB trials ($Z=-3.06$, $p_{adj(3)}= .006$, $r=0.600$) and marginally significant for the possible location of the UB trials ($Z=-1.89$, $p_{adj(3)}= .059$, $r=0.371$), suggesting that – as one could expect – participants found it generally more difficult to detect changes that occurred peripherally than those that occurred centrally, independent of belief and location type.

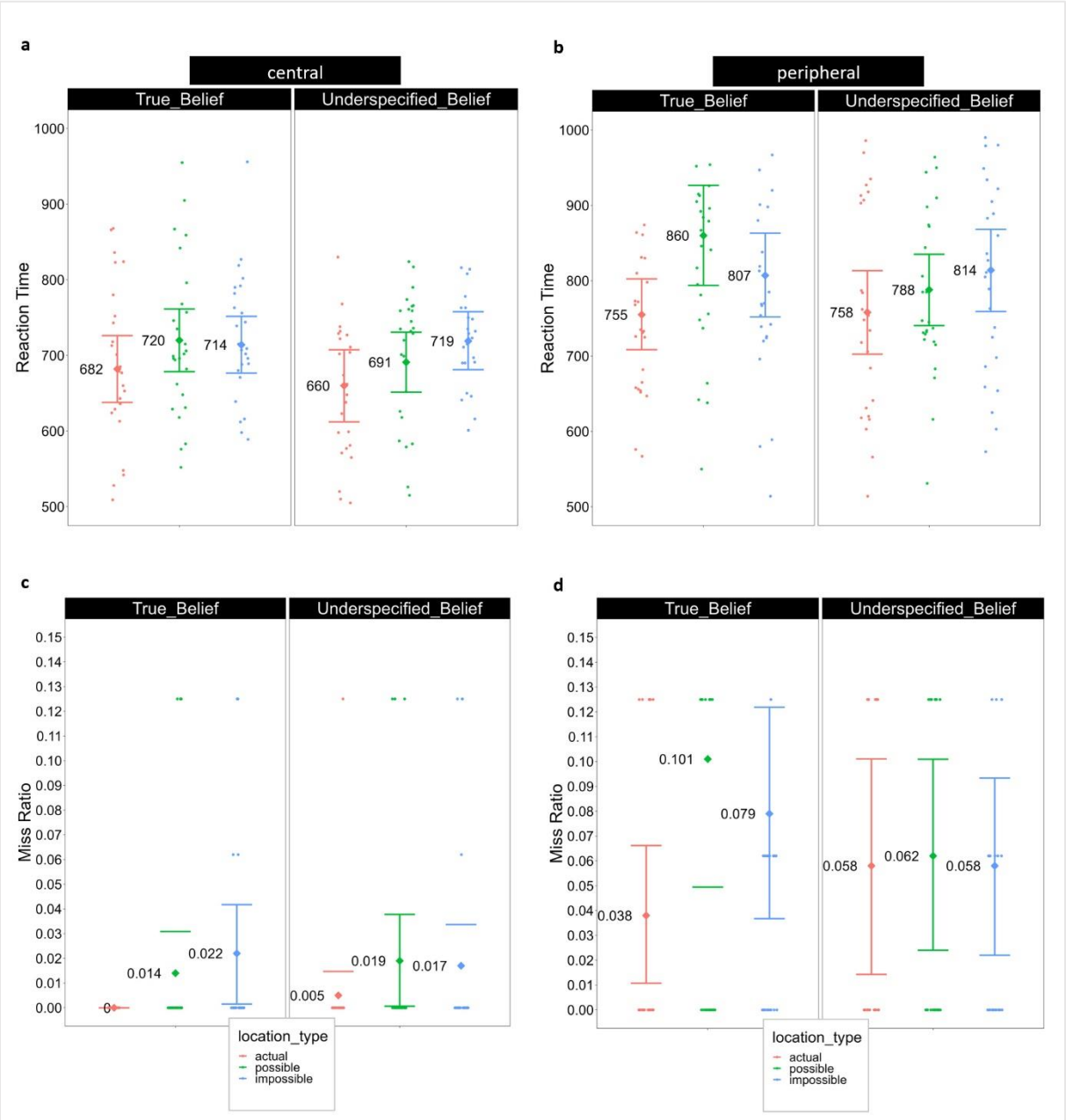


Figure S3.1. Mean reaction time (a,b) and miss ratio (c,d) for changes occurring at central (left) and peripheral (right) positions at the three types of locations, on the true and the underspecified belief trials in Experiment 1a. Error bars represent 95% CI, dots show the individual data. Note: The statistical tests for the RT differences were run on the log-transformed RT data. Only the results of the two crucial comparisons are presented (actual versus impossible and possible versus impossible) on the figures. +: p<0.1; *: p<0.05, **: p<0.01

S3.2.2 Experiment 1b

Reaction times

Results of a 2 x 2 x 3 ANOVA, position (central vs peripheral), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors revealed a significant main effect of location type $F(2, 64)=11.85$, $p< .001$, $\eta_p^2= .20$) but no significant main effect of belief ($F(1, 32)=1.85$, $p= .184$, $\eta_p^2= .055$) or belief x location type interaction ($F(2, 64)=0.15$, $p= .860$, $\eta_p^2= .005$). There was, however, a significant main effect of position ($F(1, 32)=100.72$, $p< .001$, $\eta_p^2= .759$): participants were much slower in detecting changes at the peripheral positions, compared to the central ones (see **Figure S3.3a** and **S3.3b)**. There was also a tendency level position x location type interaction ($F(2, 64)=3.03$, $p= .055$, $\eta_p^2= .086$), resulting from the fact that participants were much faster in detecting changes not only at the actual but also at the 'possible' location, i.e. on the large box that was empty, when it occupied a more central spatial position, compared to changes at the 'impossible' locations occupying the same spatial position, independent of the agent's belief, while the pattern was rather the opposite for changes occurring at peripheral spatial locations. Importantly, however, neither the position x belief ($F(1, 32)=0.13$, $p= .724$, $\eta_p^2= .004$) nor the position x belief x location type interaction ($F(1, 32)=0.17$, $p= .840$, $\eta_p^2= .005$) was significant, reflecting this similar pattern on the two types of belief trials.

In line with this, follow-up 2x3 ANOVA, run on RTs for changes occurring at central spatial positions, revealed a significant effect of location type ($F(2, 64)=12.33$, $p< .001$, $\eta_p^2= .278$), but no significant effect of belief ($F(1, 32)=2.16$, $p= .152$, $\eta_p^2= .063$) or belief x location type interaction ($F(2, 64)=0.31$, $p= .738$, $\eta_p^2= .009$). Pairwise comparisons indicated a significant difference between the possible and the impossible location on the TB ($t(32)=-2.33$, $p_{adj(3)}= .032$, $d=0.405$) but only a tendency level difference between these two types of locations the UB trials ($t(32)=-1.85$, $p_{adj(3)}= .080$, $d=0.321$). The difference between the actual and impossible location was significant for both types of trials (TB: $t(32)=-3.90$, $p_{adj(3)}< .001$, $d=0.679$; UB: $t(32)=-3.65$, $p_{adj(3)}= .003$, $d=0.636$). For peripheral changes, a follow-up 2x3 ANOVA yielded a significant effect of location type ($F(2, 64)=4.78$, $p= .012$, $\eta_p^2= .130$), but no significant effect of belief ($F(1, 32)=0.31$, $p= .584$, $\eta_p^2= .009$) or belief x location type interaction ($F(2, 64)=0.12$, $p= .886$, $\eta_p^2= .004$). Pairwise comparisons revealed significant difference only between the actual and the impossible location and only on the UB trials, after correcting for multiple comparisons ($t(32)=-4.62$, $p_{adj(3)}< .001$, $d=0.160$; all other pairwise comparisons: $ts <2.25$, $p_{unadjs}> .105$).

Altogether, the pattern of results indicates that, in Experiment 1b, participants' attention was mainly directed by the spatial position and the surface features (i.e. the size) of the boxes rather than what

the agent considers possible, with central large boxes capturing more and peripheral, smaller boxes less attention.

Miss ratios

As can be seen on **Figure S3.3c** and **S3.3d**, just like in Experiment 1a, the observed pattern of findings resulted predominantly from participants' performance on the 'peripheral'-change trials, with statistical tests indicating a significant difference between the three types of locations, specifically a significant difference between the actual and the possible and a tendency level difference between the actual and the impossible location, in the number of misses on the TB (Friedman test: $\chi^2(2)=7.36$, $p= .025$, Kendall's $W=0.112$; follow-up Wilcoxon Signed Rank Tests: actual-possible: $Z=-2.57$, $p_{adj(3)}= .030$, $r=0.397$; actual-impossible: $Z=-2.00$, $p_{adj(3)}= .092$, $r=0.397$) but not on the UB trials (Friedman test: $\chi^2(2)=1.50$, $p= .472$, Kendall's $W=0.023$). For changes at the central spatial positions, the pattern was similar for two beliefs: almost no miss for changes at the actual versus a few in case of changes at the other two locations. The difference between the three types of locations was again marginally significant for the TB ($\chi^2(2)=4.88$, $p= .087$, Kendall's $W=0.074$; follow-up Wilcoxon Signed Rank Tests: all Zs< 1.62, all p_{adj}s> .110) but not for the UB trials ($\chi^2(2)=2.00$, $p= .368$, Kendall's $W=0.033$). As in Experiment 1a, miss ratios were much higher for changes at the peripheral locations compared to the central ones, with the differences being significant for all three types of locations, independent of belief (TB – actual: $Z=-2.32$, $p_{adj(3)}= .020$, $r=0.404$; possible: $Z=-3.38$, $p_{adj(3)}= .002$, $r=0.589$; impossible: $Z=-3.69$, $p_{adj(3)}< .001$, $r=0.643$; UB – actual: $Z=-2.52$, $p_{adj(3)}= .014$, $r=0.439$; possible: $Z=-2.68$, $p_{adj(3)}= .014$, $r=0.467$; impossible: $Z=-3.34$, $p_{adj(3)}= .003$, $r=0.582$). In sum, results suggest that while participants directed more attention to the actual location when the agent's knowledge state was the same as their own, particularly when the two large boxes occupied a central spatial position, i.e. this did not require much effort, no such 'actual bias' was present on those trials, where the other agent was uncertain about the ball's hiding location, independent of the exact spatial position of the two large empty boxes.

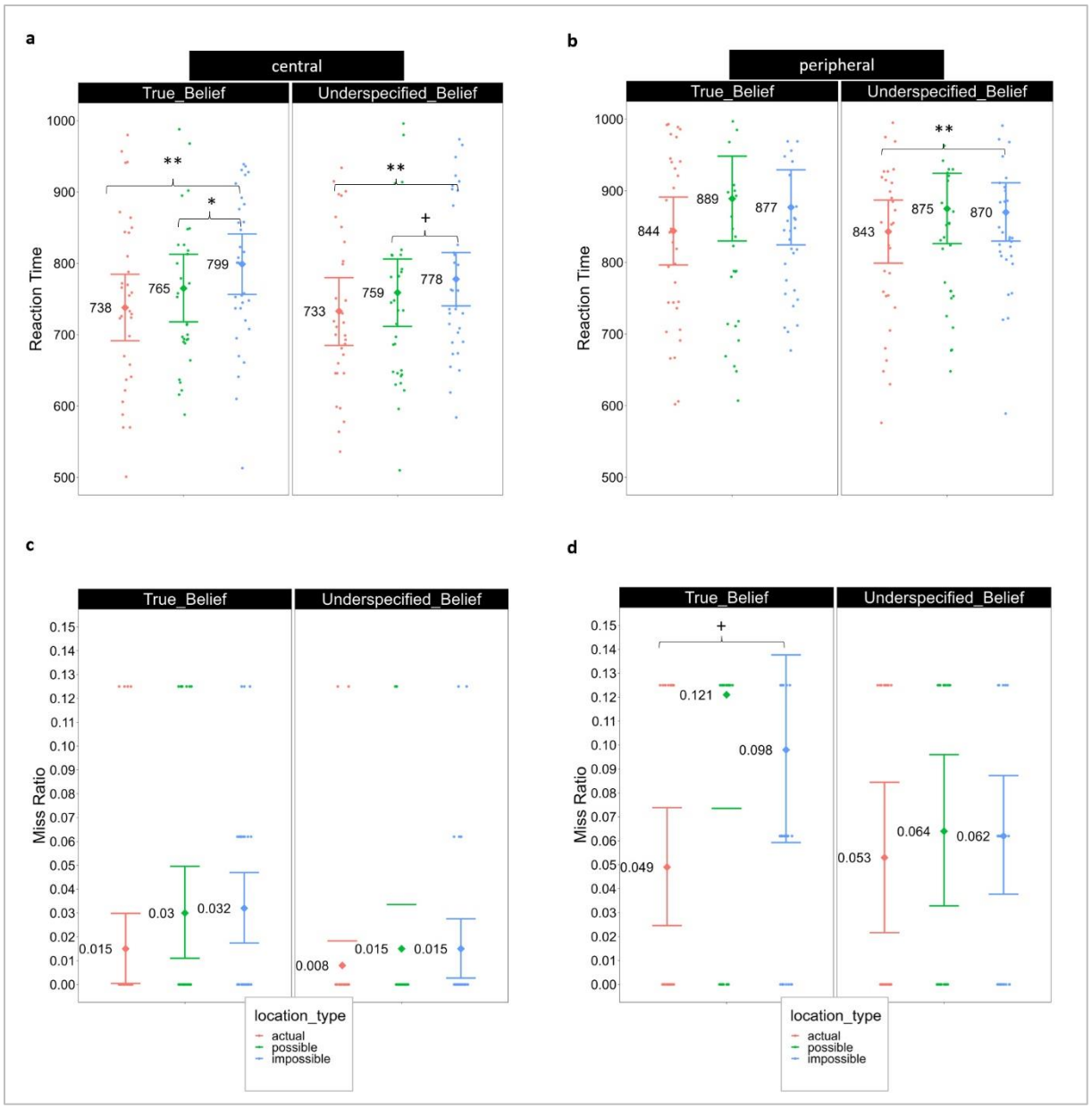


Figure S3.3. Mean reaction time (a,b) and miss ratio (c,d) for changes occurring at central (left) and peripheral (right) positions at the three types of locations, on the true and the underspecified belief trials in Experiment 1b. Error bars represent 95% CI, dots show the individual data. Note: The statistical tests for the RT differences were run on the log-transformed RT data. Only the results of the two crucial comparisons are presented (actual versus impossible and possible versus impossible) on the figures. +: p<0.1; *: p<0.05, **: p<0.01

S3.2.3 Experiment 2

Reaction times

Results of a 2 x 2 x 3 ANOVA, with position (central vs peripheral), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors, yielded a significant main effect of location

type $F(2, 68)=14.64$, $p< .001$, $\eta_p^2= .301$) but no significant main effect of belief ($F(1, 34)= 0.14$, $p= .713$, $\eta_p^2= .004$) or belief x location type interaction ($F(2, 68)=1.52$, $p= .225$, $\eta_p^2= .043$). Crucially, the analysis revealed a significant main effect of position ($F(1, 34)=177.63$, $p< .001$, $\eta_p^2= .839$): participants were much slower in detecting changes at the peripheral positions, compared to the central ones, in general (see **Figure S3.3a** and **S3.3b)**. There was no significant position x belief ($F(1, 34)=0.002$, $p= .965$, $\eta_p^2= .000$) or position x location type interaction either ($F(2, 68)=1.52$, $p= .225$, $\eta_p^2= .043$). Importantly, however, there was a significant position x belief x location type interaction ($F(2, 68)=3.17$, $p= .048$, $\eta_p^2= .085$). As can be seen on the figure, on trials where the change occurred at one of the central spatial positions, RTs were shorter for changes not only at the actual but also at the possible (i.e. at the large empty box) compared to the impossible locations (i.e. at the small boxes) on both the true and the underspecified belief trials. When the change occurred peripherally, such a difference between the possible and the impossible location was present only when the agent represented two alternatives about the ball's location (with the reality bias being present independent of belief).

A 2x3 ANOVA, run separately for changes occurring at central locations, yielded no significant main effect of belief ($F(1, 34)=0.14$, $p= .709$, $\eta_p^2= .004$). There was, however, a significant main effect of location type ($F(2, 68)=13.57$, $p< .001$, $\eta_p^2= .285$) and a significant belief x location type interaction ($F(2, 68)=4.66$, $p= .013$, $\eta_p^2= .120$). Pairwise comparisons indicated a significant difference between the possible and the impossible location on the UB ($t(34)=-3.57$, $p_{adj(3)}= .003$, $d=0.603$) and a marginally significant difference on the TB trials ($t(34)=-1.96$, $p_{adj(3)}= .058$, $d=0.331$) as well as a significant difference between the actual and the impossible location on the TB ($t(34)=-5.29$, $p_{adj(3)}< .001$, $d=0.894$) and a marginally significant on the UB trials ($t(34)=-2.16$, $p_{adj(3)}= .076$, $d=0.365$), with the pattern suggesting the lack of actual/reality bias in the underspecified belief condition. For changes occurring peripherally, a follow-up 2x3 ANOVA yielded a significant main effect of location type ($F(2, 68)=8.60$, $p< .001$, $\eta_p^2= .202$) but no significant main effect of belief ($F(1, 34)=0.52$, $p= .821$, $\eta_p^2= .004$) or belief x location type interaction ($F(2, 68)=1.29$, $p= .283$, $\eta_p^2= .036$). Pairwise comparisons revealed that the difference between the possible and the impossible location was not significant in either of the two belief conditions (TB – possible-impossible: $t(34)=-0.59$, $p_{adj(3)}= .561$, $d=0.100$; UB – possible-impossible: $t(34)=-1.08$, $p_{adj(3)}= .288$, $d=0.182$) and the difference between the actual and the impossible location was significant only in the underspecified ($t(32)=-4.24$, $p_{adj(3)}< .001$, $d=0.716$) but not in the true belief condition ($t(34)=-1.43$, $p_{adj(3)}= .241$, $d=0.242$). Altogether, these findings clearly show that the predicted effect (bias towards the possible location on the underspecified belief trials) was not simply the result of the actual spatial position occupied by the 'possible' box.

Miss ratios

As can be seen on **Figure S3.3c** and **S3.3d** pattern of findings observed in the main analyses resulted, again, mainly from participants' performance on the 'peripheral'-change trials. Friedman test indicated a significant difference between the three types of locations in terms of the number of misses on the TB trials ($\chi^2(2)=8.73$, $p= .013$, Kendall's $W=0.125$), resulting from significantly fewer number of misses in case of changes at the actual compared to other two types of locations (actual-possible: $Z=-2.77$, $p_{adj(3)}= .015$, $r=0.468$; actual-impossible: $Z=-2.36$, $p_{adj(3)}= .036$, $r=0.399$) but not on the UB trials ($\chi^2(2)=4.29$, $p= .117$, Kendall's $W=0.061$), even though number of misses was somewhat higher on these trials for changes at the possible than in the other two locations. For changes occurring centrally, Friedman tests revealed no significant difference between the three types of locations on the TB trials ($\chi^2(2)=4.27$, $p= .118$, Kendall's $W=0.061$), despite the pattern of miss ratios was similar to the one observed on peripheral-change trials. However, it indicated a significant difference between the three types of locations on the UB trials, with respect to how many times participants failed to detect changes at those ($\chi^2(2)=10.55$, $p= .015$, Kendall's $W=0.151$). Post hoc Wilcoxon Signed Rank revealed significant difference between the possible and the impossible ($Z=-2.92$, $p_{adj(3)}= .012$, $r=0.493$) and tendency level difference between the possible and the actual location ($Z=-2.00$, $p_{adj(3)}= .092$, $r=0.338$): participants had basically no misses on those UB trials where the change occurred at the possible location, when the respective box occupied a central spatial position.

Just like in the previous experiments, miss ratios were much higher for changes at the peripheral positions compared to the central ones, with the differences being significant in all experimental conditions, both on the true (actual: $Z=-4.60$, $p_{adj(3)}< .001$, $r=0.777$; possible: $Z=-4.16$, $p_{adj(3)}< .001$, $r=0.703$; impossible: $Z=-4.05$, $p_{adj(3)}< .001$, $r=0.684$) and on the underspecified belief trials (actual: $Z=-3.09$, $p_{adj(3)}= .020$, $r=0.522$; possible: $Z=-4.72$, $p_{adj(3)}= .002$, $r=0.797$; impossible: $Z=-4.68$, $p_{adj(3)}< .001$, $r=0.790$).

Taken together, while these findings indicate a clear attentional bias towards the possible location on those trials where the other agent represented two equally likely alternatives regarding the location of the ball and when changes occurred at central spatial positions, corroborating our findings with the reaction times measure, this bias was much less clear for peripheral-changes, when monitoring the 'possible' location required more effort from the participants (as shorter RTs were accompanied by more misses).

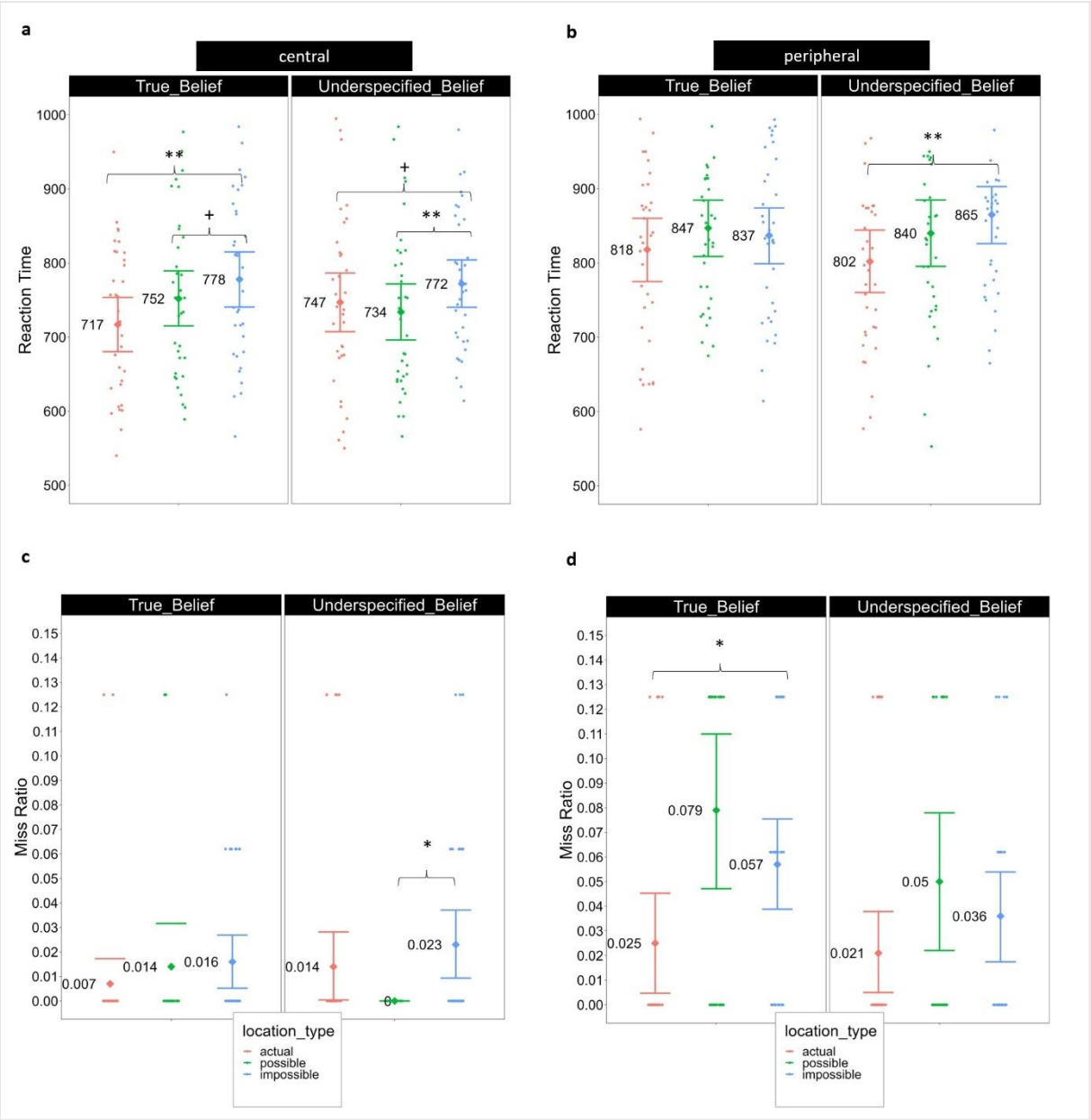


Figure S3.3. Mean reaction time (a,b) and miss ratio (c,d) for changes occurring at central (left) and peripheral (right) positions at the three types of locations, on the true and the underspecified belief trials in Experiment 2. Error bars represent 95% CI, dots show the individual differences. The statistical tests for the RT differences were run on the log-transformed RT data. Only the results of the two crucial comparisons are presented (actual versus impossible and possible versus impossible) on the figures. +: p<0.1; *: p<0.05, **: p<0.01

S3.2.4 Experiment 3

Reaction times

Results of a 2 x 2 x 3 ANOVA, position (central vs peripheral), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors revealed a significant main effect of condition

($F(2, 70)=17.96$, $p< .001$, $\eta_p^2= .339$), a tendency level main effect of belief ($F(1, 35)=3.33$, $p= .077$, $\eta_p^2= .087$) but no belief x location type interaction ($F(1, 35)=0.13$, $p= 879$, $\eta_p^2= .004$). There was, however, a significant main effect of position ($F(1, 35)=148.22$, $p< .001$, $\eta_p^2= .809$): as can be seen on **Figure S3.4a** and **S3.4b** participants were much slower in detecting changes at the peripheral positions, compared to the central ones. There was also a tendency level position x location type interaction ($F(1.57, 54.95)=2.64$, $p= .093$, $\eta_p^2= .070$). Participants were faster in detecting changes not only at the actual but also at the 'possible' location, when it occupied a central spatial position (compared to changes at the 'impossible' locations occupying the same spatial position), on both types of belief trials, while there was either no such difference between the possible and the impossible location (TB condition) or there was an opposite pattern (longer RTs for changes at the possible location in the UB condition) on the peripheral-change trials. Despite the somewhat different reaction time pattern in the true and the underspecified belief condition, in those cases when the change occurred peripherally, neither the position x belief ($F(1, 35)=0.76$, $p= .389$, $\eta_p^2= .021$) nor the position x belief x location type interaction ($F(2, 70)=0.39$, $p= .679$, $\eta_p^2= .011$) was significant.

A follow-up 2x3 ANOVA, run on RTs for changes occurring centrally, revealed a significant main effect of location type ($F(2, 70)=25.73$, $p< .001$, $\eta_p^2= .424$) but no significant main effect of belief ($F(1, 35)=0.95$, $p= .337$, $\eta_p^2= .026$) or belief x location type interaction ($F(2, 70)=0.09$, $p= .910$, $\eta_p^2= .003$). Pairwise comparisons indicated significant difference between all three types of locations on both the TB (actual-possible: $t(35)=-2.72$, $p_{adj(3)}= .012$, $d=0.452$; actual-impossible: $t(35)=-5.20$, $p_{adj(3)}< .001$, $d=0.867$; possible-impossible: $t(35)=-2.90$, $p_{adj(3)}= .006$, $d=0.483$) and the UB trials (actual-possible: $t(35)=-2.58$, $p_{adj(3)}= .022$, $d=0.430$; actual-impossible: $t(32)=-5.19$, $p_{adj(3)}< .001$, $d=0.465$; possible-impossible: $t(35)=-2.69$, $p_{adj(3)}= .022$, $d=0.447$). Follow-up 2x3 repeated-measures ANOVA run on peripheral changes, yielded a significant main effect of location type ($F(1.62, 56.85)=5.61$, $p= .006$, $\eta_p^2= .138$) but no significant main effect of belief ($F(1, 35)=2.74$, $p= .107$, $\eta_p^2= .073$) or belief x location type interaction ($F(2, 70)=0.32$, $p= .728$, $\eta_p^2= .009$). Pairwise comparisons revealed significant difference between the actual and the impossible location ($t(35)=-2.87$, $p_{adj(3)}= .021$, $d=0.478$) and a tendency level difference between the actual and the possible location ($t(35)=-2.13$, $p_{adj(3)}= .080$, $d=0.355$), but no significant difference between the possible and the impossible condition on the TB trials ($t(35)=-0.14$, $p_{adj(3)}= .890$, $d=0.023$). There was no significant difference between the actual and the other two types of locations on the UB trials, after correcting for multiple comparisons (actual-possible: $t(35)=-2.07$, $p_{adj(3)}= .192$, $d=0.345$; actual-impossible: $t(35)=-1.71$, $p_{adj(3)}= .138$, $d=0.285$), neither between the possible and the impossible locations ($t(35)=0.80$, $p_{adj(3)}= .431$, $d=0.133$).

Altogether, these results indicate that, in Experiment 3 (just like in Experiment 1b before), it was mainly the spatial position and the size of the boxes that determined the direction of participants' attention

rather than which location the agent considered a potential hiding place for the object, with much less attention directed at peripheral than central, large boxes.

Miss ratios

As can be seen on **Figure S3.4c** and **S3.4d** on the 'peripheral-change' trials, participants demonstrated a clear, though nonsignificant, attentional bias towards the actual location of the ball, in the true belief condition (Friedman test: $\chi^2(2)=5.54$, $p= .063$, Kendall's $W=0.077$; follow-up Wilcoxon Signed Rank Tests: actual-possible: $Z=-1.93$, $p_{adj(3)}= .159$, $r=0.322$; actual-impossible: $Z=-1.91$, $p_{adj(3)}= .159$, $r=0.318$), with no significant difference between the possible and the impossible location in the number of misses (possible-impossible: $Z=-.12$, $p_{adj(3)}= .908$, $r=0.020$). Despite the similar pattern, on the underspecified belief trials the difference between the three types of locations was not significant (Friedman test: $\chi^2(2)=3.98$, $p= .137$, Kendall's $W=0.055$). Surprisingly, on trials where changes occurred centrally, participants missed fewer changes at the possible than at the other two types of locations, independent of the avatar's belief. The difference between the three types of locations was, however, significant only for the UB ($\chi^2(2)=13.73$, $p= .001$, Kendall's $W=0.191$) but not for the TB trials ($\chi^2(2)=3.64$, $p= .162$, Kendall's $W=0.051$), with follow-up Wilcoxon Signed Rank Tests indicating a significant difference only between the possible and the impossible ($Z=-3.07$, $p_{adj(3)}= .006$, $r=0.512$) but not between the possible and the actual location on the UB trials ($Z=-0.82$, $p_{adj(3)}= .414$, $r=0.137$). There was no difference between the actual and the impossible location either, after adjusting for multiple comparisons ($Z=-1.76$, $p_{adj(3)}= .156$, $r=0.293$). Just like in the previous experiments, miss ratios were much higher for changes occurring at peripheral compared to changes occurring at central locations. The differences were significant for all three types of locations, except for the actual location on the TB trials (TB – actual: $Z=-1.36$, $p_{adj(3)}= .175$, $r=0.227$; possible: $Z=-3.26$, $p_{adj(3)}= .003$, $r=0.543$; impossible: $Z=-3.30$, $p_{adj(3)}$ .003, $r=0.550$; UB – actual: $Z=-2.24$, $p_{adj(3)}= .025$, $r=0.373$; possible: $Z=-3.13$, $p_{adj(3)}= .006$, $r=0.522$; impossible: $Z=-2.89$, $p_{adj(3)}= .008$, $r=0.382$). In sum, the general pattern of findings corroborates our reaction time results, indicating a clear influence of the ball's actual location on participants' attention, with the possible bias being most likely the result of proximity of the ball's actual hiding place to the possible location on the central-change trials, rather than the content of the agent's belief in these cases.

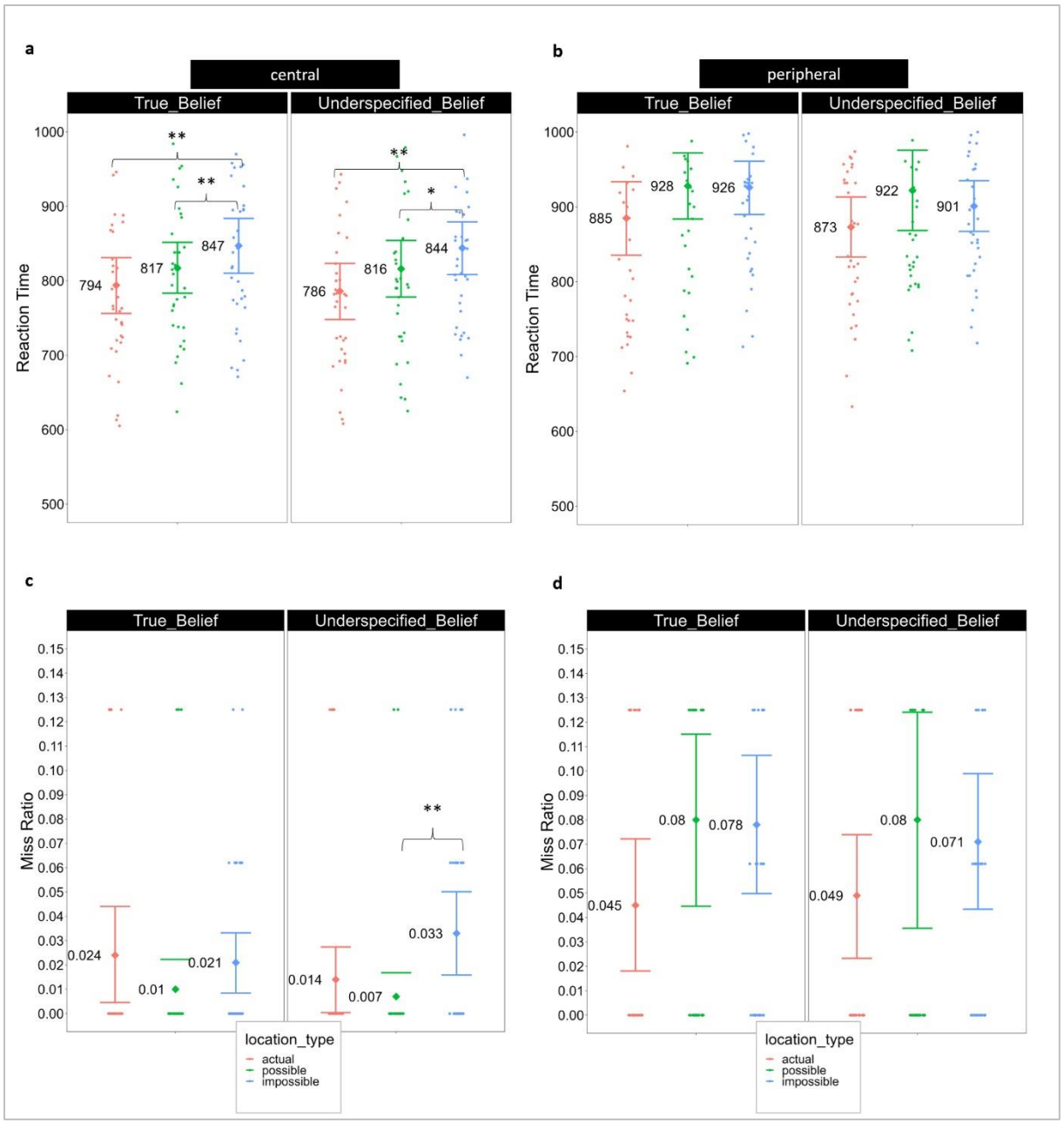


Figure S3.4. Mean reaction time (a,b) and miss ratio (c,d) for changes occurring at central (left) and peripheral (right) positions at the three types of locations, on the true and the underspecified belief trials in Experiment 3. Error bars represent 95% CI and dots show the individual data. Note: The statistical tests for the RT differences were run on the log-transformed RT data. Only the results of the two crucial comparisons are presented (actual versus impossible and possible versus impossible) on the figures. +: p<0.1; *: p<0.05, **: p<0.01

S3.2.5 Experiment 4 – ball present trials

Reaction Times

Results of a 2 x 2 x 3 ANOVA, with position (central vs peripheral), belief (TB versus UB) and location type (actual, possible, impossible) as within-subject factors as within-subject factors revealed a significant main effect of location type ($F(2, 68)=25.38$, $p< .001$, $\eta_p^2= .427$) and belief ($F(1, 34)=4.64$, $p= .038$, $\eta_p^2= .120$) but no significant belief x location type interaction ($F(1.54, 52.37)=0.36$, $p= .933$, $\eta_p^2= .001$). The main effect of position was, again, significant ($F(1, 34)=69.13$, $p< .001$, $\eta_p^2= .671$), reflecting participants' longer RTs for changes at the peripheral positions (compared to the central ones), see **Figure S3.5a** and **S3.5b**. There was also a tendency level position x location type interaction ($F(2, 68)=3.05$, $p= .054$, $\eta_p^2= .082$): participants were much faster in detecting changes at the actual location (compared to changes at the 'impossible' locations), when the respective box occupied a central spatial position, than when it occupied peripheral one, independent of the belief of the agent. Crucially, however, neither the position x belief ($F(1, 34)=0.55$, $p= .465$, $\eta_p^2= .016$) nor the position x belief x location type interaction ($F(1.64, 55.89)=0.68$, $p= .481$, $\eta_p^2= .020$) was significant. Follow-up 2x3 ANOVA, run on RTs for changes occurring at central spatial positions, indicated a significant main effect of belief ($F(1, 34)=5.54$, $p= .025$, $\eta_p^2= .140$) and location type ($F(2, 68)=27.65$, $p< .001$, $\eta_p^2= .448$), but no significant belief x location type interaction ($F(1.73, 58.73)=0.44$, $p= .617$, $\eta_p^2= .013$). Pairwise comparisons revealed significant difference between the actual and the other two types of locations on both type of belief trials (TB - actual-impossible: $t(34)=-4.26$, $p_{adj(3)}< .001$, $d=0.720$, actual-possible: $t(34)=-2.3$, $p_{adj(3)}= .050$, $d=0.395$; UB -actual-impossible: $t(34)=-5.86$, $p_{adj(3)}< .001$, $d=0.990$; actual-possible: $t(34)=-4.06$, $p_{adj(3)}< .001$, $d=0.686$) and a marginally significant difference between the possible and the impossible location on the UB ($t(34)=-2.01$, $p_{adj(3)}= .052$, $d=0.340$) but not on the TB trials ($t(34)=-1.63$, $p_{adj(3)}= .111$, $d=0.276$). Although the pattern was clearly the same on the 'peripheral-change' trials, follow-up 2x3 ANOVA yielded significant main effect for the location type only ($F(2, 68)=4.76$, $p= .012$, $\eta_p^2= .123$), but no significant main effect of belief ($F(1, 34)=1.09$, $p= .304$, $\eta_p^2= .031$) or belief x location type interaction ($F(1.51, 51.25)=0.34$, $p= .654$, $\eta_p^2= .010$). Despite the lack of significant interaction, pairwise comparisons indicated only a marginally significant difference, and only between the actual and the impossible location, on the TB trials, after correcting for multiple comparisons ($t(34)=-2.43$, $p_{adj(3)}= .060$, $d=0.411$; all other comparisons: ts <1.74, p_{adj}s> .183). Altogether, these results show that the effects observed in our main analyses, i.e. the difference between participants' RTs for changes at the possible and the impossible location, were independent of the actual spatial position occupied by the 'possible' and the 'impossible' box.

Miss ratios

Friedman tests, run separately for the central- and peripheral-change trials, indicated no significant difference between the three types of locations in terms of the number of misses, on either the true (central: $\chi^2(2)=2.44$, $p= .296$, Kendall's $W=0.035$; peripheral: $\chi^2(2)=2.00$, $p= .368$, Kendall's $W=0.029$) or the underspecified belief trials (central: $\chi^2(2)= 2.91$, $p= .234$, Kendall's $W=0.042$; peripheral: $\chi^2(2)=2.19$, $p= .334$, Kendall's $W=0.031$).

Just like in the previous experiments, miss ratios were higher for changes occurring at peripheral compared to changes occurring at central locations. The differences were significant for the TB possible ($Z=-2.67$, $p_{adj(3)}= .024$, $r=0.451$) and UB impossible trials ($Z=-3.17$, $p_{adj(3)}= .006$, $r=0.535$) and marginally significant for the TB impossible trials ($Z=-2.14$, $p_{adj(3)}= .066$, $r=0.361$).

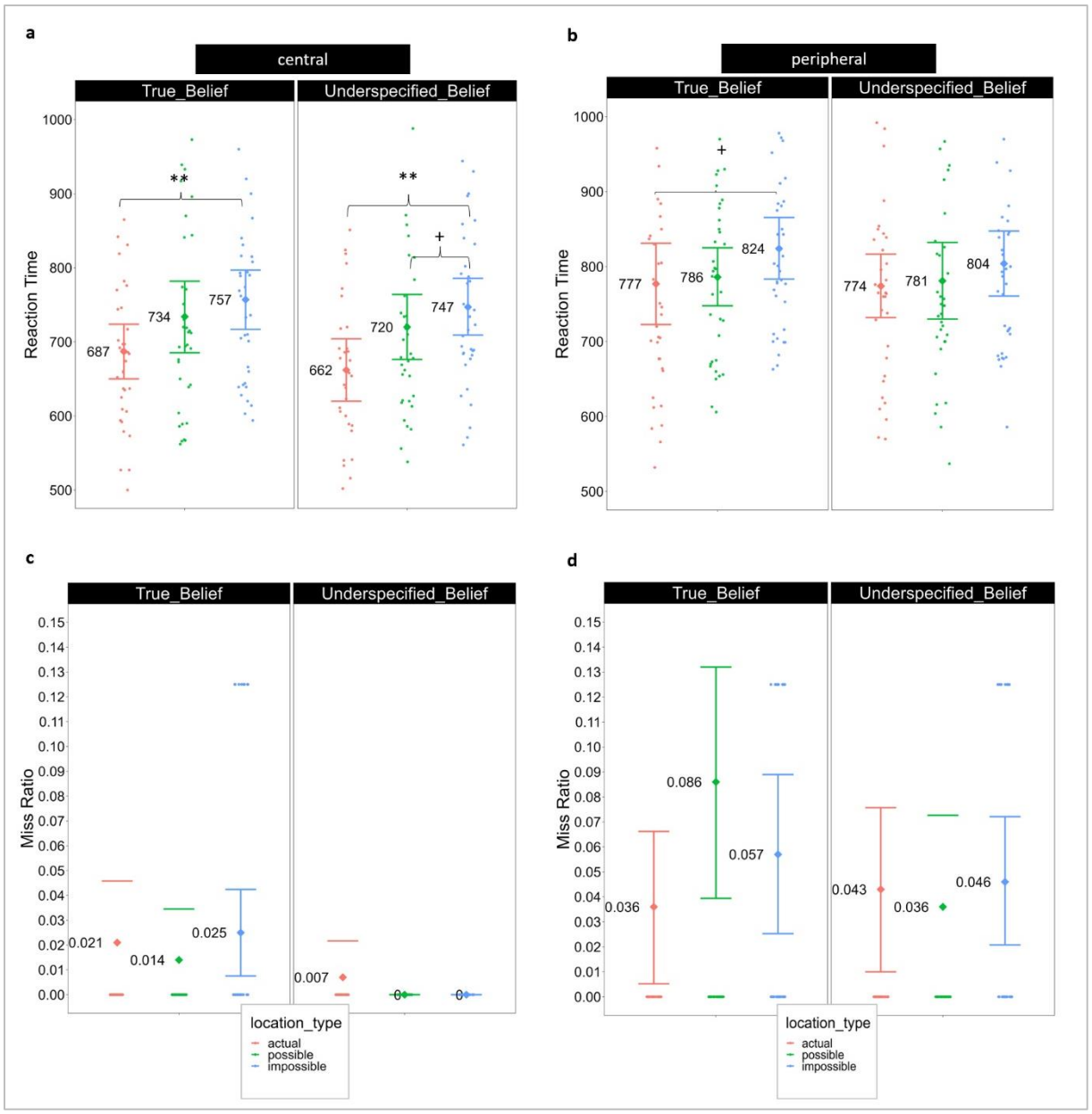


Figure S3.5. Mean reaction time (a,b) and miss ratio (c,d) for changes occurring at central (left) and peripheral (right) positions at the three types of locations, on the 'ball present' true and the underspecified belief trials of Experiment 4. Error bars represent 95% CI, dots show the individual data. Note: The statistical tests for the RT differences were run on the log-transformed RT data. Only the results of the two crucial comparisons are presented (actual versus impossible and possible versus impossible) on the figures. +: p<0.1; *: p<0.05, **: p<0.01

S3.2.6. Experiment 4 – ball absent trials

Reaction Times

Results of a 2 x 2 x 2 repeated measures ANOVA, with position (central vs peripheral), belief (TB versus UB) and location type (possible, impossible) as within-subject factors, revealed that the observed effects did not depend on the actual spatial position occupied by the 'possible' box. There was a significant main effect of belief ($F(1, 34)=13.22$, $p< .001$, $\eta_p^2= .280$) and a tendency level main effect of location type ($F(1, 34)=2.88$, $p= .099$, $\eta_p^2= .078$), but no significant belief x location type interaction ($F(1, 34)=0.56$, $p= .461$, $\eta_p^2= .016$). As can be seen on **Figure S3.6a** and **S3.6b**, although participants were generally slower in detecting changes at the peripheral positions, compared to the central ones (position main effect: $F(1, 34)=82.55$, $p< .001$, $\eta_p^2= .708$), the pattern of RT results was the same for the two spatial locations, with neither the position x location type ($F(1, 34)=0.75$, $p= .394$, $\eta_p^2= .021$), nor position x belief ($F(1, 34)=0.003$, $p= 956$, $\eta_p^2= .000$) or the position x belief x location type interaction being significant ($F(1, 34)=0.01$, $p= .933$, $\eta_p^2= .000$).

Miss ratios

Wilcoxon Signed Rank Tests (run separately for the two positions and the true and underspecified belief trials) revealed no significant difference between the possible and the impossible location in terms of the number of misses, on either the central (TB: $Z=-0.00$, $p_{adj(2)}=1.00$; UB: $Z=-0.82$, $p_{adj(2)}= .712$, $r=0.139$) or the peripheral-change trials (TB: $Z=-1.41$, $p_{adj(2)}= .316$, $r=0.238$; UB: $Z=-0.92$, $p_{adj(2)}= .712$, $r=0.162$), corroborating the RT results which indicated no effect of spatial position on participants' change detection performance. Except for the UB possible trials, just like in Experiment 1a-3, miss ratios were higher for changes occurring at peripheral locations than for changes occurring at more central spatial positions (see **Figure S3.6c** and **S3.6d**), indicating difficulties with detecting peripheral changes. The difference was significant for the true belief possible ($Z=-3.27$, $p= .002$, $r=0.552$) and impossible ($Z=-2.84$, $p= .005$, $r=0.480$) and marginally significant for the UB impossible trials ($Z=-2.16$, $p= .066$, $r=0.364$) but not for the UB possible trials ($Z=-0.71$, $p= .480$, $r=0.120$).

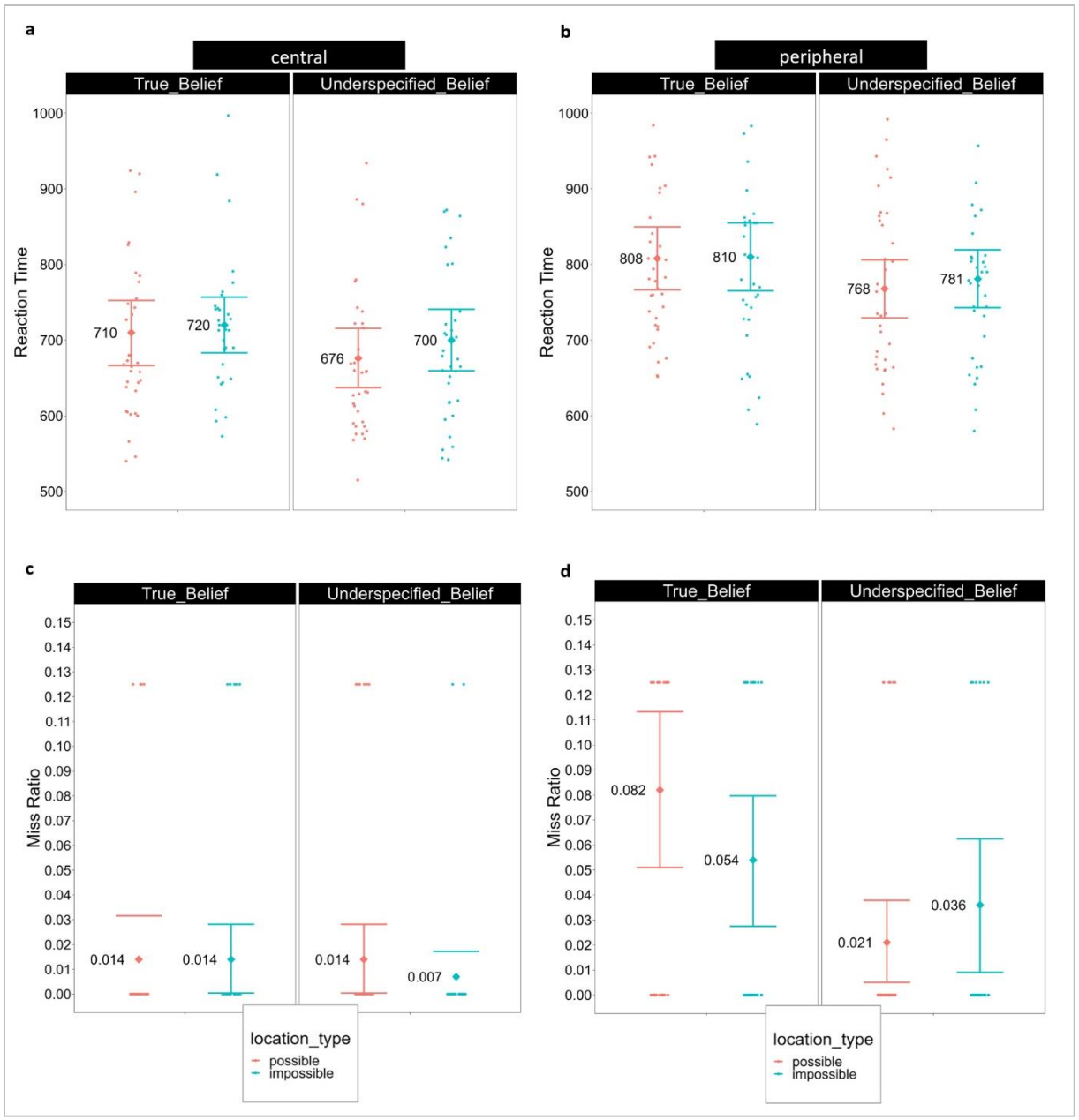


Figure S3.6. Mean reaction time (a,b) and miss ratio (c,d) for changes occurring at central (left) and peripheral (right) positions at the two types of locations, on the 'ball absent' true and the underspecified belief trials of Experiment 4. Error bars represent 95% CI, dots show individual data. Note: As interactions with time were not significant, no follow-up tests were run on the (log-transformed) RT data. +: p<0.1; *: p<0.05, **: p<001

S3.3 Individual heterogeneity in the five experiments

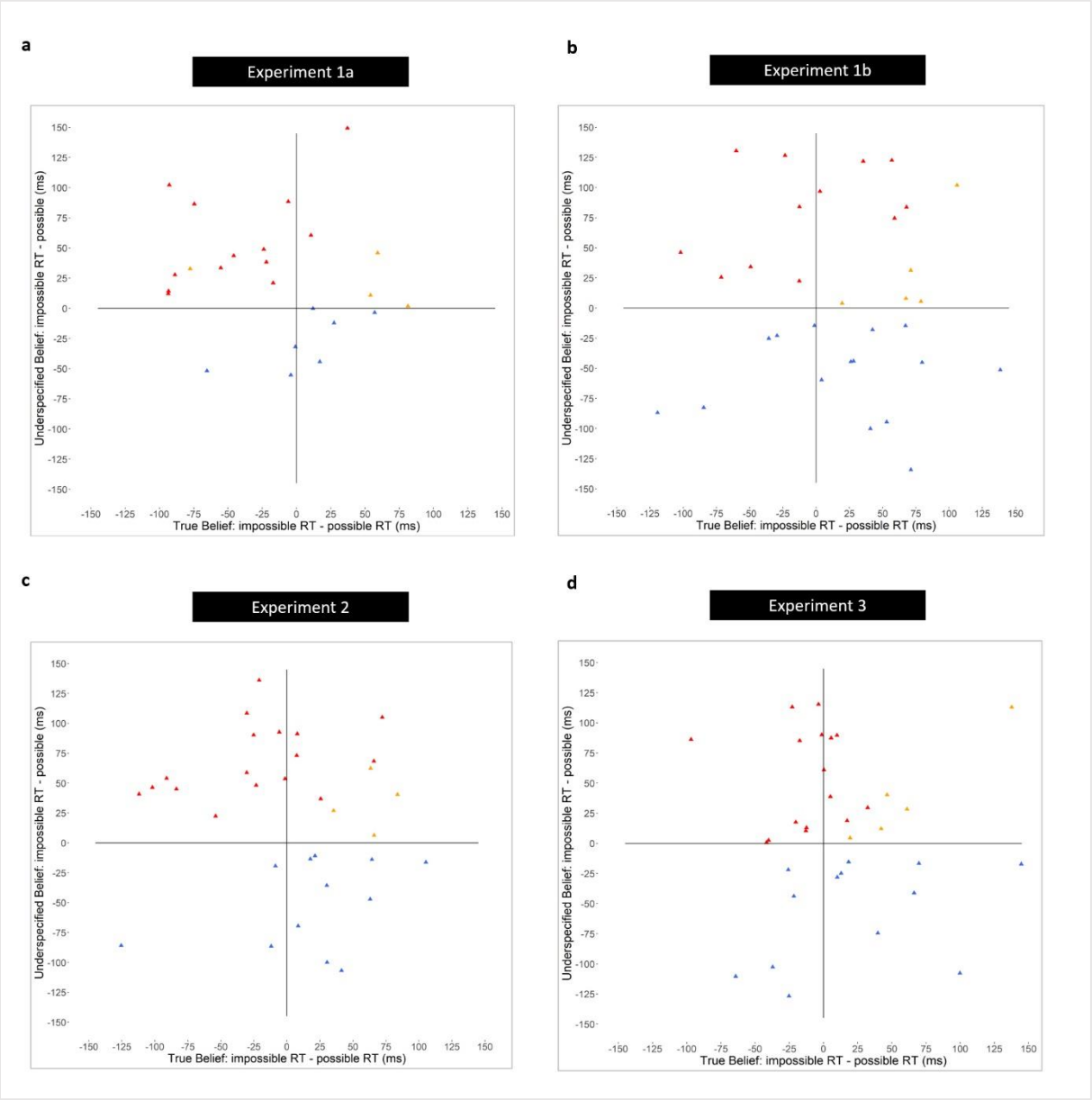


Figure S3.7. UB difference score (Y-axis) as the function of TB difference score (X-axis) in Experiment 1a-3. Positive difference scores mean that the participant detected changes faster at the possible (but empty) location than at the impossible locations. Red: participants whose difference scores were in line with our predictions (UB difference score was positive and the TB difference score was either negative or positive but smaller in magnitude than the corresponding UB difference score), i.e. who may have represented the two alternatives from third-person perspective. Orange: participants for whom the predicted effect was present on the UB trials but who may have just represented the alternatives from first-person perspective (as the effect was not only present but even larger on the TB trials). Blue: Participants for whom the predicted effect was not present.

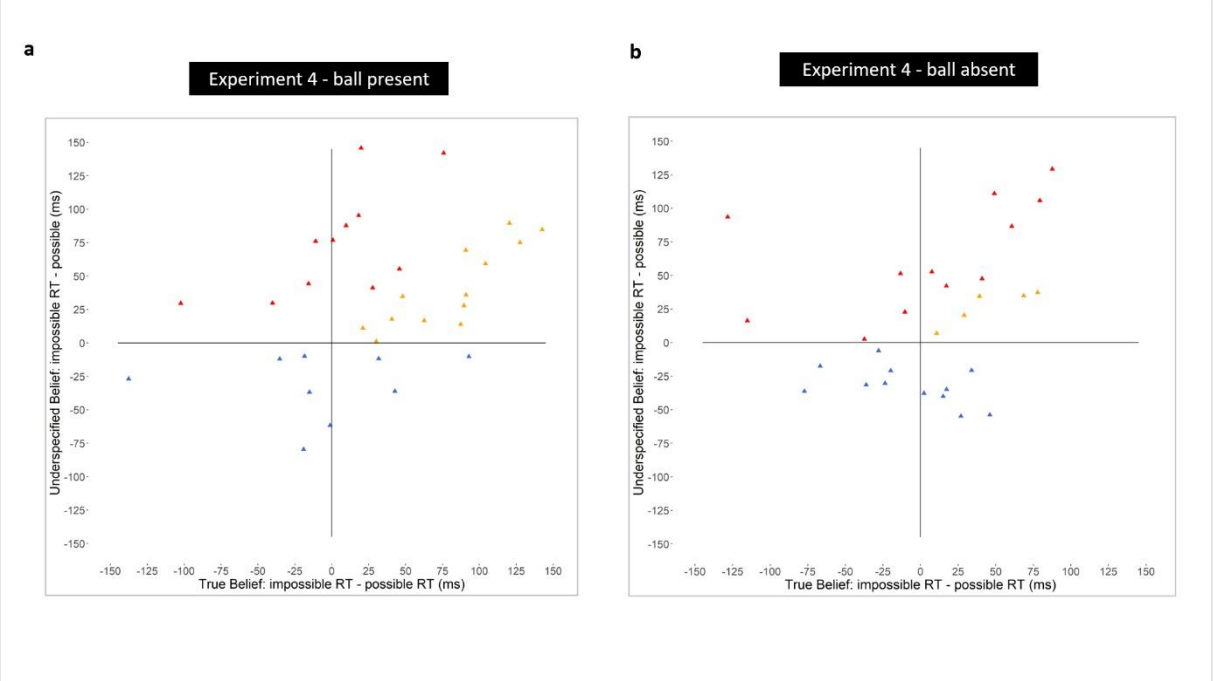


Figure S3.8. UB difference score (Y-axis) as the function of TB difference score (X-axis) in (a) the 'ball present' and (b) the 'ball absent' trials of Experiment 4. Red: participants whose UB and TB difference scores were in line with the predictions. Orange: participants for whom the predicted effect was present on the UB trials but who may have just represented the alternatives from first-person perspective (as the effect was not only present but even larger on the TB trials). Blue: Participants for whom the predicted effect was not present.

S3.4 Comparison of the first and the second half of the task (the effect of 'time'): figures

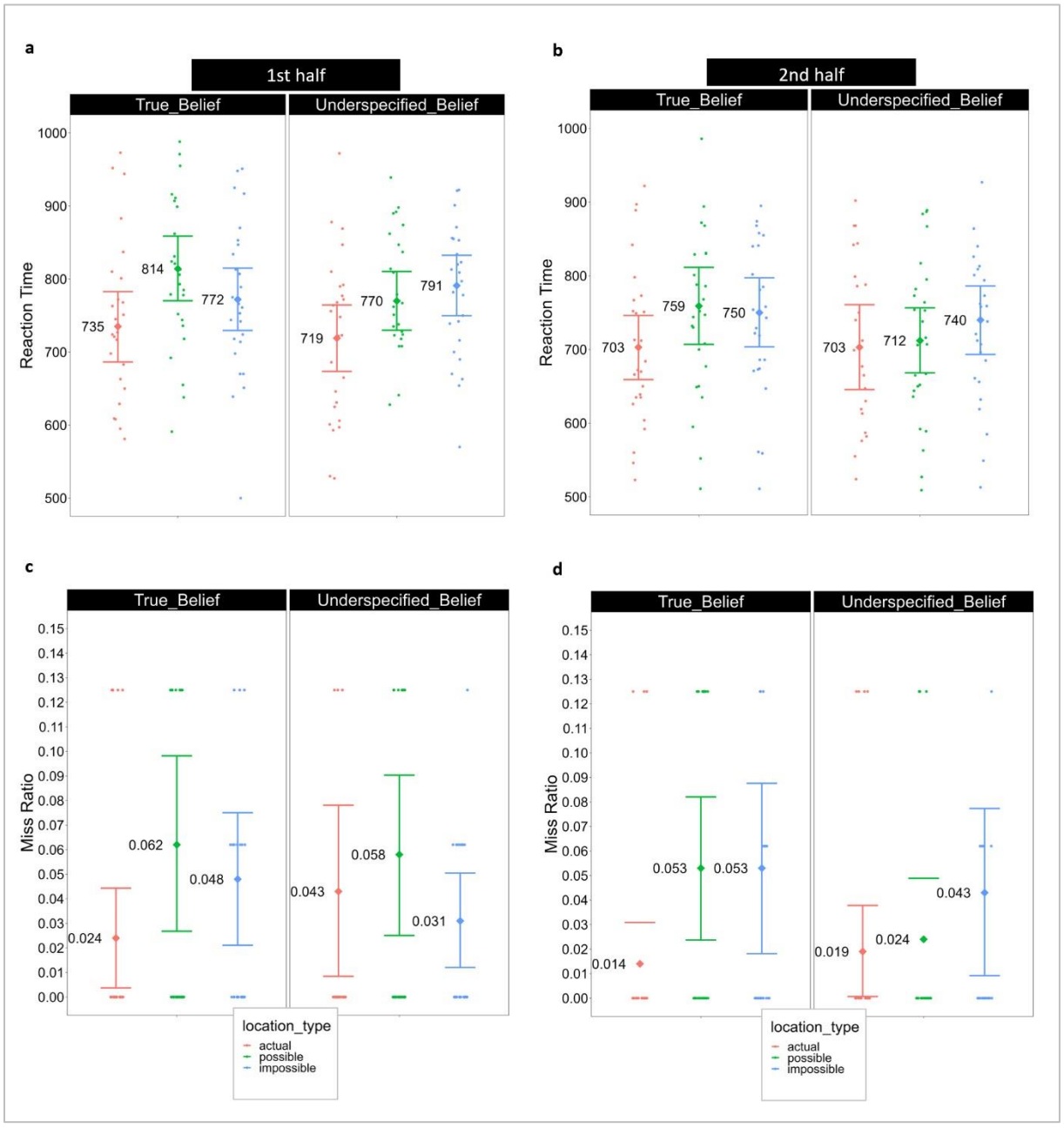


Figure S3.9. Mean reaction time (a,b) and miss ratio (c,d) at the three types of locations, on the true and the underspecified belief trials in the first (left) and the second half of the experiment (right) in Experiment 1a. Error bars represent 95% CI, dots show individual data. Note: The figure serves only an illustrative purpose. Since neither the interactions with time (RT) nor the Friedman tests (miss ratio) were significant, no follow-up tests were run on the data.

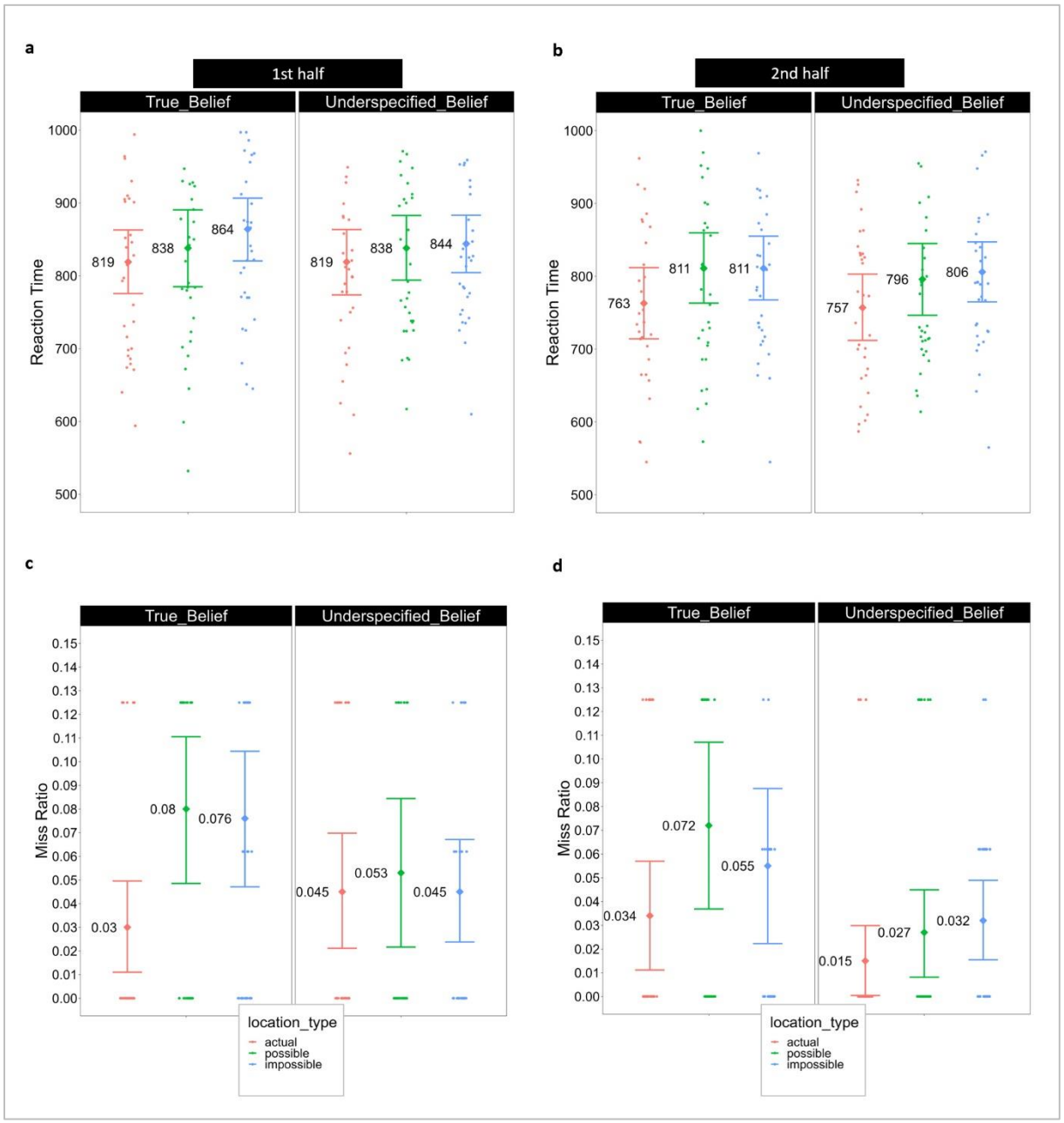


Figure S3.10. Mean reaction time (a,b) and miss ratio (c,d) at the three types of locations, on the true and the underspecified belief trials in the first (left) and the second half of the trials (right) in Experiment 1b. Error bars represent 95% CI. Dots show the individual data. Note: The figure serves only an illustrative purpose. Since neither the interactions with time (RT) nor the Friedman tests (miss ratio) were significant, no follow-up tests were run on the data.

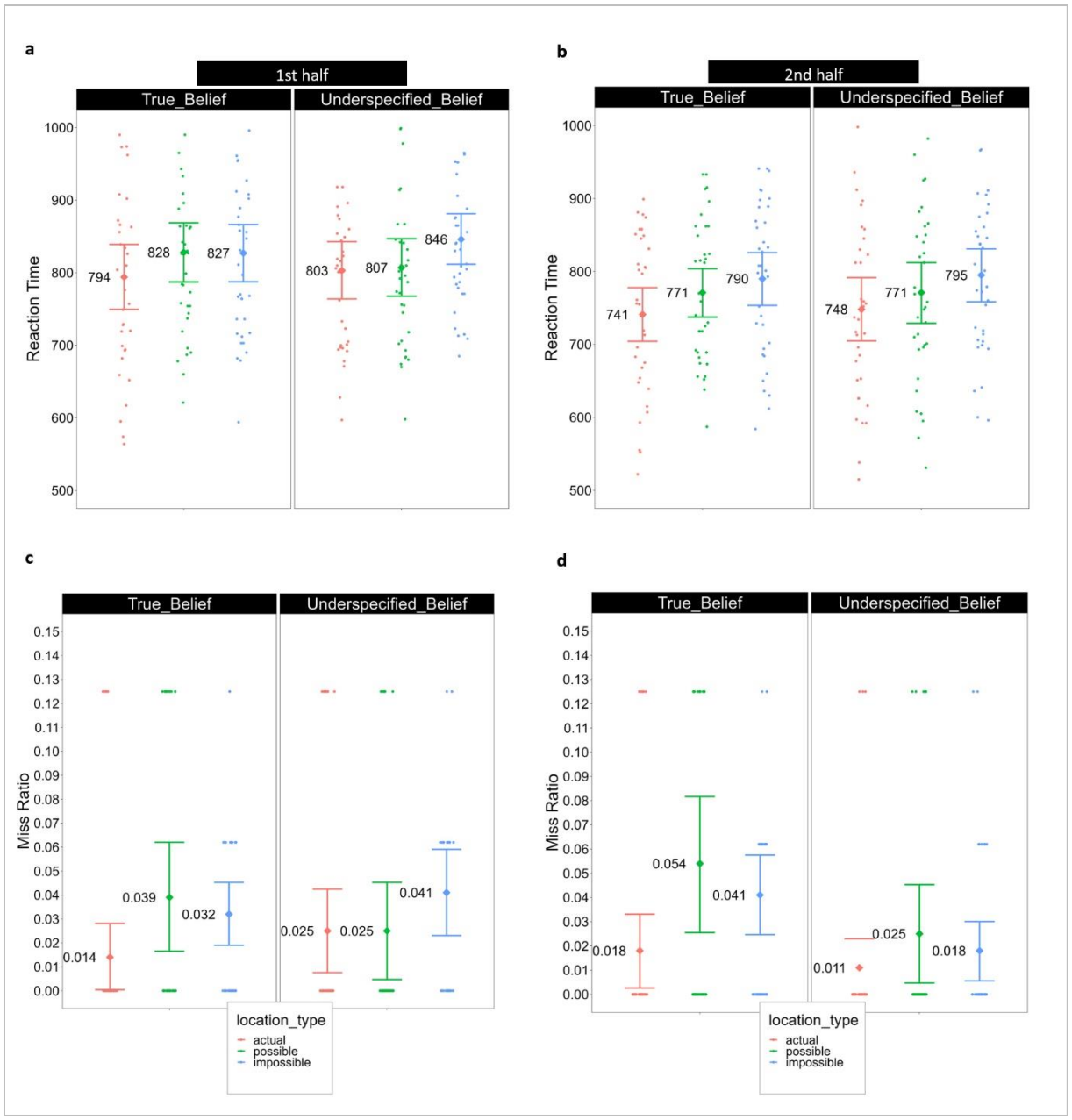


Figure S3.11. Mean reaction time (a,b) and miss ratio (c,d) at the three types of locations, on the true and the underspecified belief trials in the first (left) and the second half of the trials (right) in Experiment 2. Error bars represent 95% CI, dots show individual data. Note: The figure serves only an illustrative purpose. Since neither the interactions with time (RT) nor the Friedman tests (miss ratio) were significant, no follow-up tests were run on the data.

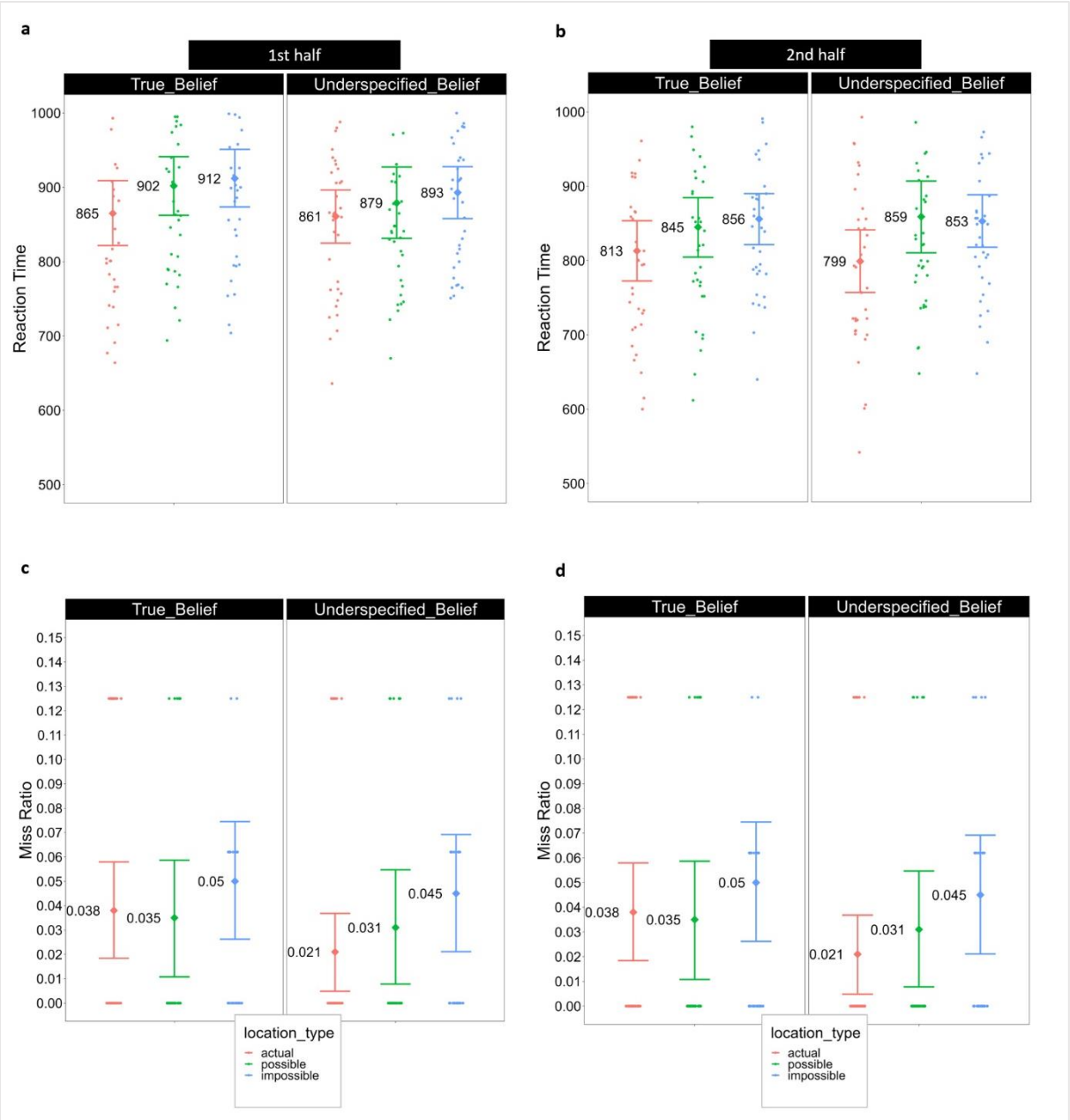


Figure S3.12. Mean reaction time (a,b) and miss ratio (c,d) at the three types of locations, on the true and the underspecified belief trials in the first (left) and the second half of the trials (right) in Experiment 3. Error bars represent 95% CI, dots show individual data. Note: The figure serves only an illustrative purpose. Since neither the interactions with time (RT) nor the Friedman tests (miss ratio) were significant, no follow-up tests were run on the data.

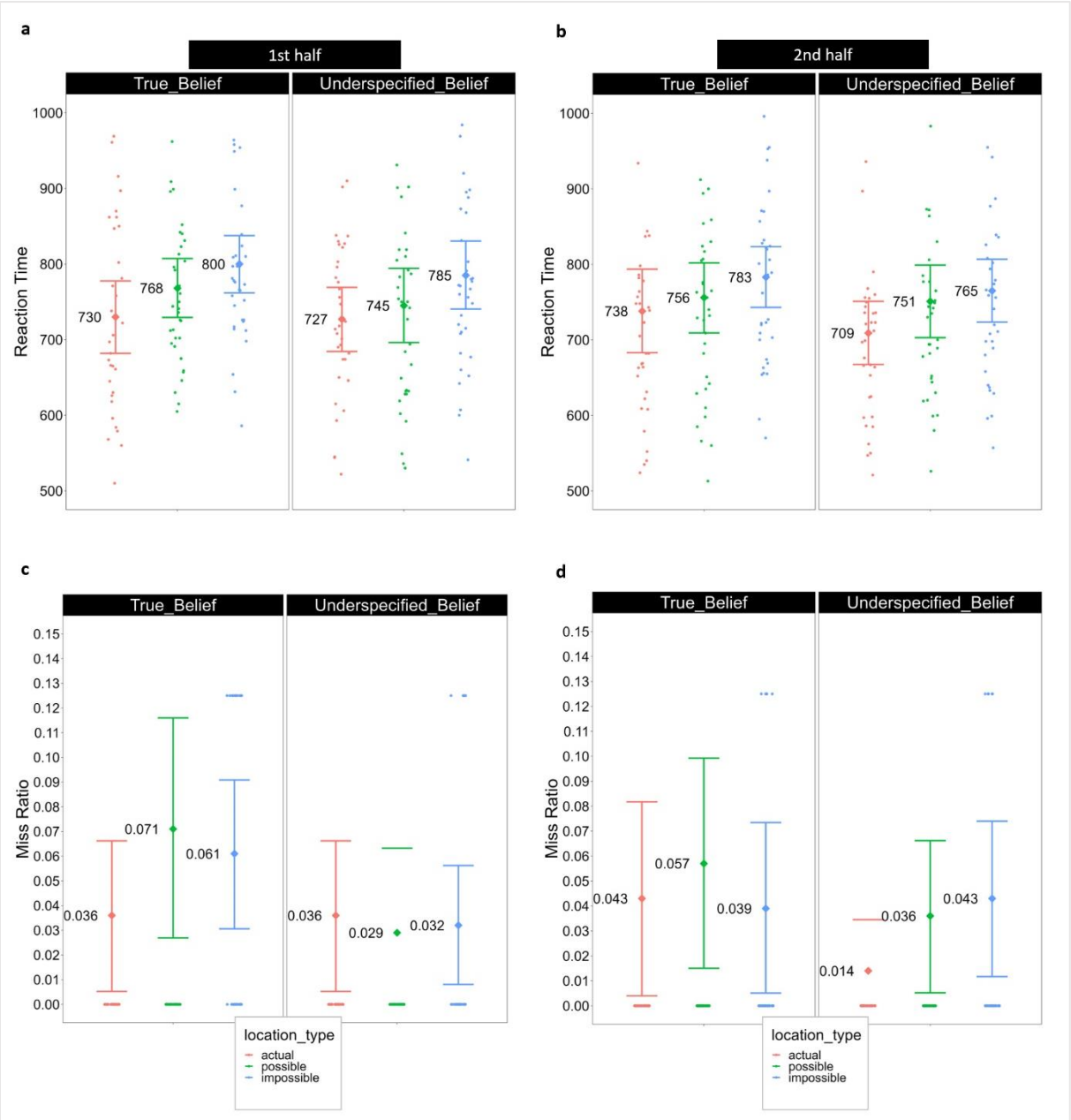


Figure S3.13. Mean reaction time (a,b) and miss ratio (c,d) at the three types of locations, on the true and the underspecified belief trials in the first (left) and the second half of the trials (right) in the 'ball present' trials of Experiment 4. Error bars represent 95% CI, dots show individual data. Note: The figure serves only an illustrative purpose. Since neither the interactions with time (RT) nor the Friedman tests (miss ratio) were significant, no follow-up tests were run on the data.

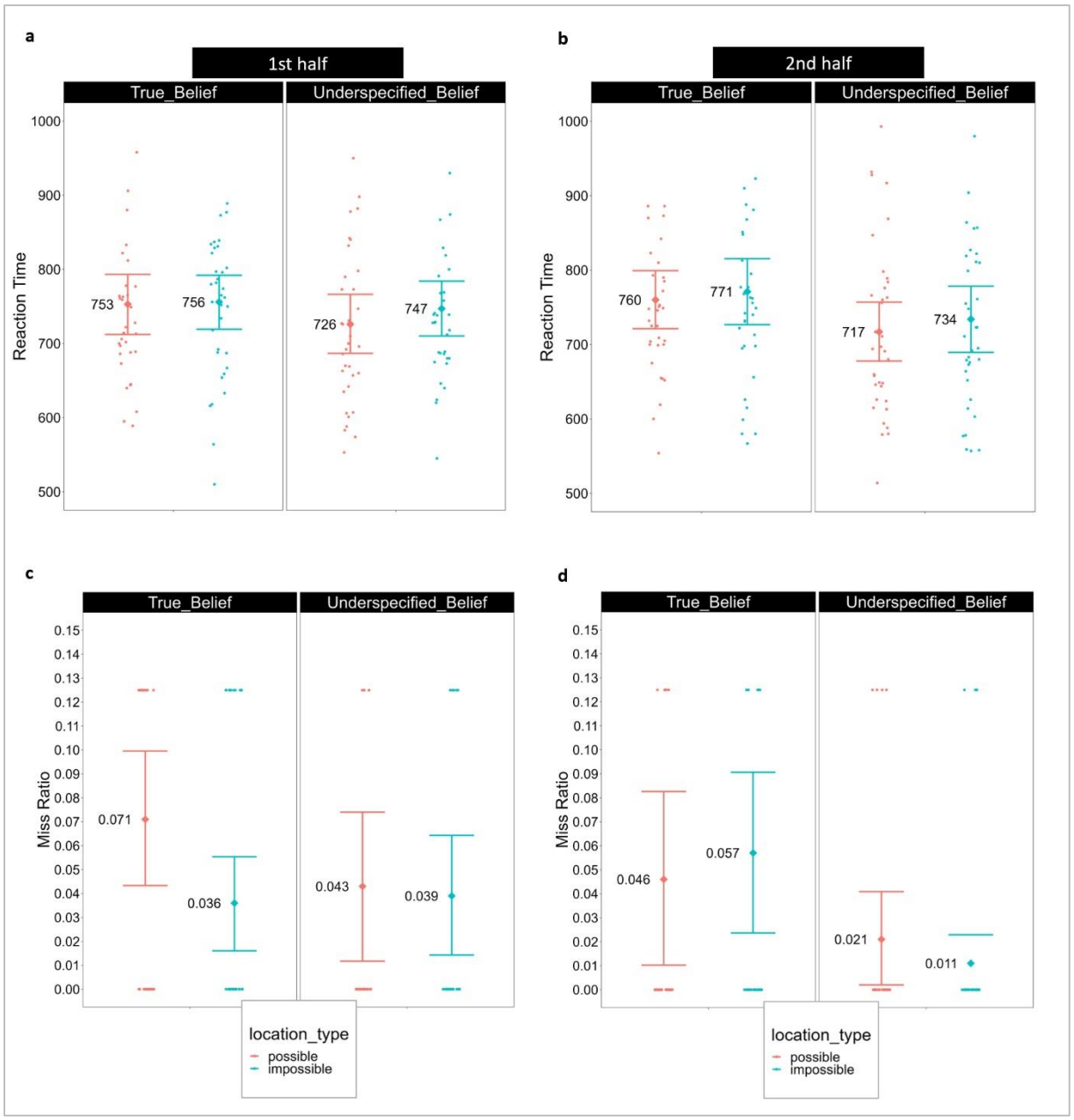


Figure S3.14. Mean reaction time (a,b) and miss ratio (c,d) in the two types of locations, on the true and the underspecified belief trials in the first (left) and the second half of the trials (right) in the 'ball absent' trials Experiment 4. Error bars represent 95% CI, dots show individual data. Note: The figure serves only an illustrative purpose. Since the interactions with time (RT) were not significant, no follow-up tests were run on the data.

S3.5 Experiment 4: main analyses of the 'ball absent' trials with three locations

Reaction times

A 2 x 3 repeated-measures ANOVA with belief (TB versus UB) and location type (possible1, possible2, impossible) as within-subject factors, yielded a significant main effect of belief (F(1, 34)=14.15, p= .001, η_p^2= .294), resulting from the fact that participants were faster on the underspecified belief trials, in general, as well as a significant main effect of location type (F(2, 68)=3.23, p= .046, η_p^2= .087). Participants were much faster to detect changes at the possible1 location (i.e. when the change occurred at the possible location that was approached first by the ball and towards which it last faced before leaving) than at the impossible location, on both the TB and the UB trials (see **Figure S3.15a)**. In case of the 'possible2' location reaction times differed markedly on the two types of belief trials: on the TB trials RTs were much higher in case the change occurred at this than in case it occurred at the other possible location, while there was no difference between the two on the UB trials (in line with our predictions). In spite of the different reaction time patterns the belief x location type interaction was not significant (F(2, 68)=1.31, p= .277, η_p^2= .037). After adjusting for multiple comparisons, paired-samples t-tests revealed a marginally significant difference between two possible locations (t(34)=-2.34, $p_{adj(3)}$= .075, d=0.396) but no significant difference between the impossible and the two possible locations on the TB trials (possible1-impossible: t(34)=-1.70, $p_{adj(3)}$= .196, d=0.287; possible2-impossible: t(34)=-0.47, $p_{adj(3)}$= .640, d=0.080) and no significant difference between the three types of locations on the UB trials (possible1-possible2: t(34)=-0.51, $p_{adj(3)}$= .959, d=0.009; possible1-impossible: t(34)=-1.65, $p_{adj(3)}$= .294, d=0.279; possible2-impossible: t(34)=-1.70, $p_{adj(3)}$= .294, d=0.287). Altogether, these results imply that while participants' spatial attention may have been directed by the content of the agent's belief in the first place on the underspecified belief trials (that he considers the two large boxes two, equally likely alternatives), on the true belief trials it may have been driven mainly by low-level factors, such as the memory trace of which 'possible' box was approached first (and last) by the ball (resulting in an attentional bias toward the respective box and/or inhibition of the other 'possible' location).

Miss ratios

Friedman test, performed on the TB trials, indicated no significant difference between the three types of locations in terms of the number of misses ($\chi^2(2)=1.00$, $p= .607$, Kendall's $W=0.014$). In contrast, there was a significant difference between the three locations on the UB trials, with respect to how many times participants failed to detect changes at those ($\chi^2(2)=9.10$, $p= .011$, Kendall's $W=0.130$). As can be seen on **Figure S3.15b**, participants missed few changes at the possible1 and the impossible and missed literally none at the possible2 location. In line with this, follow-up Wilcoxon Signed Tests revealed significant difference between the possible2 and the other two (possible2-possible1: $Z=-2.89$, $p_{adj(3)}= .012$, $r=0.488$; possible2-impossible: $Z=-2.59$, $p_{adj(3)}= .020$, $r=0.438$) but not between the possible1 and the impossible location ($Z=-1.11$, $p_{adj(3)}= .266$, $r=0.187$).

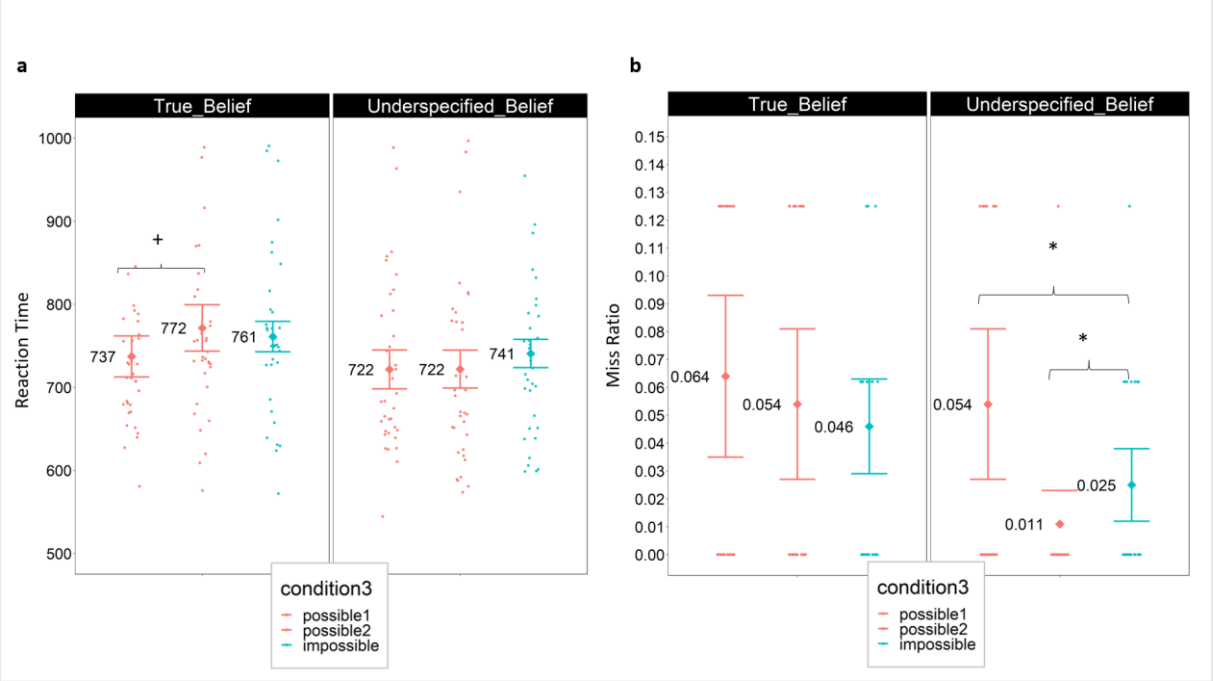


Figure S3.15. (a) Mean reaction time and (b) miss ratio in the true and underspecified belief conditions of the 'ball absent' trials, for changes occurring at the possible1, possible2 and impossible locations. Error bars represent 95% CI, dots show the individual data. Possible1: the large box (potential hiding location) first approached by the ball and towards which it last faced. Possible2: the box towards which it moved second. Note: Only the results of the two crucial comparisons are presented (actual versus impossible and possible versus impossible) on the figures. +: p<0.1; *: p<0.05, **: p<001

Supplementary Materials for Chapter 4

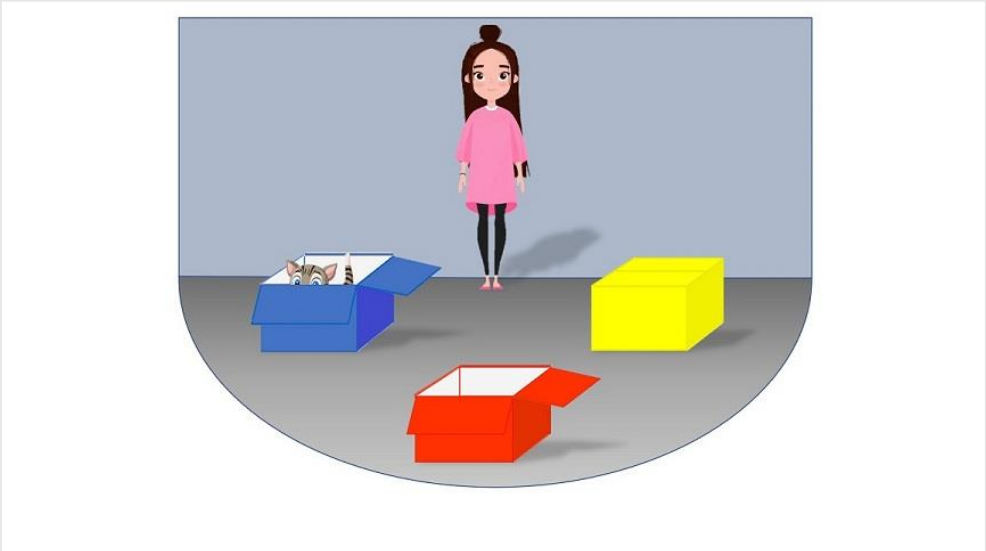
S4.2.1. Experiment 1: Instructions

General instruction

You will participate in an experiment that will have three parts. First you will just make simple decisions about pictures by clicking on them. This part will last for about one minute. Following this you will proceed to the second part, lasting for 2-3 minutes, then to the third, main part of the task. You will receive detailed instructions and a short practice session prior the main task. Sometimes progress will be self-paced but there will be a maximum time set within which you will have to proceed. You have a maximum of 40 minutes to finish the whole experiment. Please click on NEXT to start the first part!

Training session 1

You will see a few pictures, like the one below, with two of three boxes open. You will have to decide whether either of the open boxes contains a kitten. If an open box contains a kitten, then you have to click on that box, very quickly. If none of the open boxes contain kitten, do not do anything, the trial will proceed after 2 seconds. You can use either mouse or touchpad for responding.



Training session 2

You will participate in a decision-making task. In each trial you will see a sequence of pictures, with the human figure and the three different boxes you have seen before, in Part I. At the beginning of the trials the kitten you saw in Part I will hide in one of the boxes. You will not see where, just the closed boxes. Then, two of the boxes will open, one after the other, with the girl either witnessing what happens or not. At the end of each trial you will be asked whether the girl would search for the kitten in ONE or TWO boxes and you will receive a feedback after your decision.

When the girl witnesses both events, she will have all the information to infer the location of the kitten, therefore she will know in which box the kitten is. When she misses one of the events (does not see the second one), she will not have all the information to perform this inference. Thus, in these situations she cannot be certain about the kitten's location and she might consider more than one box a potential hiding location.

Please click on NEXT or wait until the task proceeds (within 3 minutes)!

Main Task (before the four practice trials)

In the following section you will see similar picture sequences as before. However, this time:

1. Each trial will begin with a word, either SHE or YOU.

2. At the end of each trial you will be asked HOW LIKELY **it is that the kitten is in a certain box (circled on the presented picture), according to YOU (in trials that start with the word YOU) or according TO THE GIRL (in trials which start with the word SHE).** You will have to indicate your response by clicking on the grey scale you saw in the calibration phase, as fast as possible! (Please click on the scale itself and not on the arrows!) In some of the trials, you will be presented with the picture of the three boxes, and you will be asked where the kitten is.

You will have to respond to this question by clicking on the box you think the kitten has hidden.

Important: Each trial will start with a fixation cross. You will have to click on the cross, where the two lines cross each other, very quickly. A trial is considered valid, only if you do this!

Now let's see the practice trials! Please click on START or wait until the task proceeds (within 3 minutes)!

Main Task (after the practice trials)

Now you will receive 2 blocks of 36 trials, similar to the ones you just saw, with no feedback. There will be one break that can last for a maximum of 3 minutes. From time to time, you will be presented with the picture of the three boxes and you will be asked where the kitten is. There you will have to click on the selected box.

Please do not forget to click on the fixation cross at the beginning of the trials! A trial is valid only if you do so!

Press START to start the main task!

S4.1.2 Experiment 2: Instructions

Training session

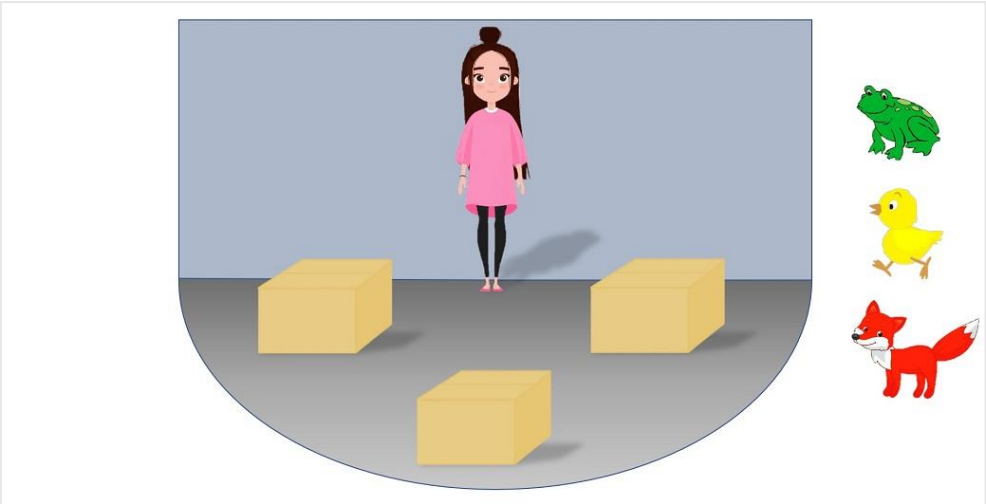
In each trial you will see a sequence of pictures, with a girl, three boxes, and the animals, presented below. At the beginning of the trials each animal will hide in a box. You will not see where, just the closed boxes on the first picture. Then, two of the boxes will open, one after the other, revealing which animal is inside, with the girl either witnessing what happens or not.

At the end of each trial, you will be presented with the picture of one of the three animals and you will be asked whether the girl would search for that animal in ONE box or TWO boxes. You will receive feedback after your decision.

Important:

When the girl witnesses both openings, she will have all the information to infer which animal is hiding in the third box, therefore she will know the location of each of the three animals. When she does not see the second opening, she will not have all the information to perform this inference. In these situations, she cannot know for sure which of the two animals is hiding in the box that opened second and which one is hiding in the third box that remained closed at the end. Therefore, she might consider more than option when looking for one of those two animals.

Please click on NEXT or wait until the task proceeds (within 3 minutes)!



S4.1.3 Experiment 3: Learning Task – Instructions

Learning phase

You will now participate in a learning task.

The animals you saw before like to hide in the boxes, but they always hide in one specific box. You will first see three trials to learn which animal hides in which box.

You do not have to do anything, just pay attention. Please press NEXT to continue.

Test phase

One of the animals has hidden in the scene.

You will now see pictures with one closed and two open boxes that will be followed by a picture of the three animals. You will have to indicate which animal has hidden in the closed box (on the previous picture), by clicking on the picture of the animal.

Please press START to continue.

S4.1.4 Experiment 4: Learning Task – Instructions

Learning Phase

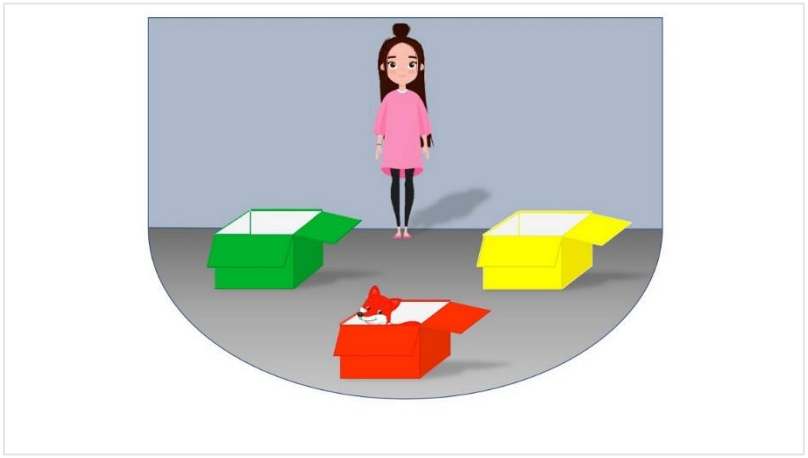
You will now participate in a learning task. The animals you saw before like to hide in the boxes.

They always hide by their colour.

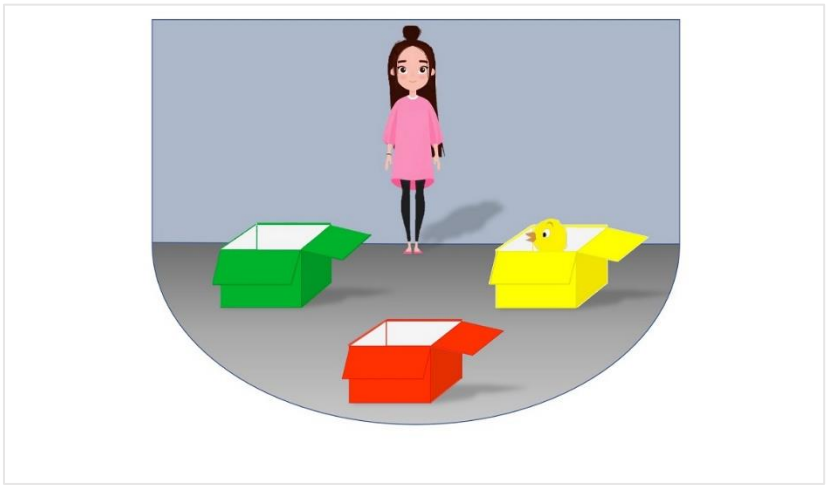
You will first see three trials to learn which animal hides in which box. You do not have to do anything, just pay attention.

Please press NEXT to continue.

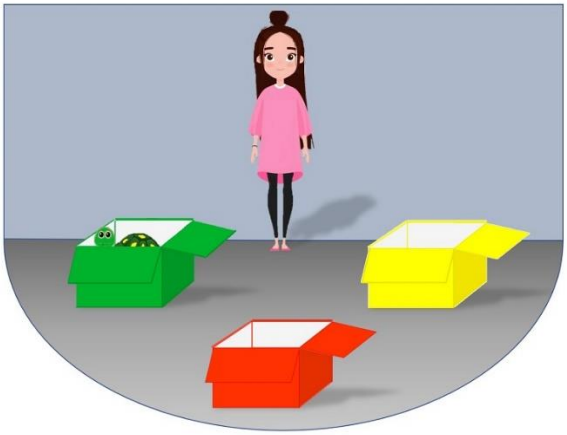
The fox can only hide in the RED box. The GIRL also knows this.

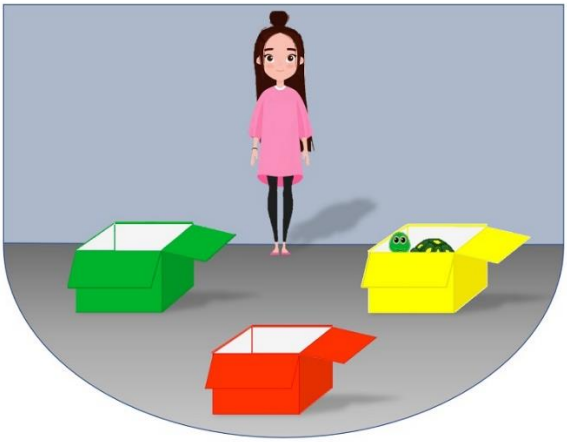


The chick can only hide in the YELLOW box. The GIRL also knows this.



The turtle can hide both in the GREEN and the YELLOW box. The GIRL also knows this.

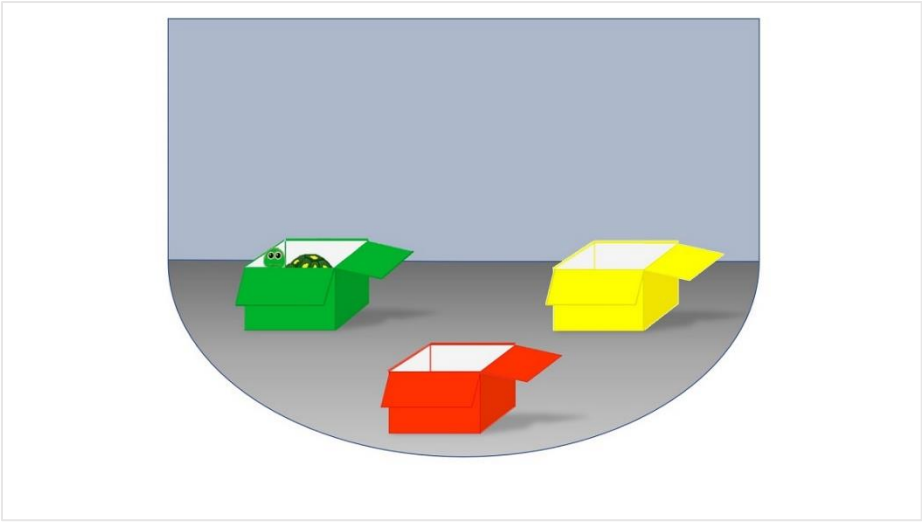




Belief correction phase

However, there is an information THE GIRL DOES NOT KNOW! The yellow patches were just painted on the turtle's back. The turtle is in fact completely green!

Because animals hide by their colour and the turtle is actually green, it can only hide in the GREEN box! The GIRL DOES NOT know this!



Test (memory) phase

One of the animals has hidden in the scene.

You will now see pictures with one closed and two open boxes that will be followed by a picture of the three animals. You will have to indicate which animal has hidden in the closed box (on the previous picture), by clicking on the picture of the animal.

After doing this, you will be always asked what response *would the girl give* to this question. You will have to reply again by clicking on the picture of the animal.

In each trial, you can click on the picture of either one or two animals (one after the other, quickly).

Remember, the girl does not know that the turtle is in fact green and therefore it can only hide in the green box!

Please press START to continue.

4.2.1 Rating data exclusion rate

Table S1. *The mean percentage of invalid trials (SD) per perspective, belief and location/animal alternative type, in Experiments 1, 2 and 3. Trials were considered invalid, and hence were not included in the analyses, if the participant (1) failed to provide response or (2) did not click on the fixation cross at the beginning of the trial (set as a requirement to ensure that the cursor appears in the middle of the scale on the response screen, at the end.)*

	belief	alternative type	Experiment 1 N=(35)	Experiment 2 (N=34)	Experiment 3 (N=35)
OTHER	True Belief	actual	4.90% (1.26%)	3.23% (7.96%)	2.08% (5.60%)
		possible	2.45% (5.99%)	2.15% (5.68%)	4.17% (8.47%)
		impossible	3.92% (9.23%)	3.23% (7.96%)	3.13% (6.61%)
	Underspecified Belief	actual	2.94% (6.45%)	3.23% (6.69%)	3.65% (7.00%)
		possible	1.96% (6.82%)	2.15% (7.13%)	4.17% (8.47%)
		impossible	1.47% (4.80%)	1.61% (5.01%)	2.08% (5.60%)
		Total	1.28%(2.94%)	2.60%(0.71%)	3.21%(0.96%)
SELF	True Belief	actual	2.94% (6.45%)	3.76% (7.08%)	4.17% (8.47%)
		possible	2.45% (7.26%)	1.61% (5.01%)	2.60% (6.15%)
		impossible	0.49% (2.86%)	0.54% (2.99%)	4.17% (9.47%)
	Underspecified Belief	actual	1.47% (4.80%)	1.61% (5.01%)	5.73% (9.09%)
		possible	1.96% (6.82%)	2.15% (7.13%)	3.13% (6.61%)
		impossible	2.45% (5.99%)	2.15% (5.68%)	2.60% (6.15%)
		Total	1.96%(0.88%)	1.97%(1.06%)	3.73%(1.21%)

Table S4.2. *The mean percentage of invalid trials (SD) per perspective, belief and alternative type, in Experiment 4.*

	belief	alternative type	Experiment 4 (N=35)
OTHER	**True Belief-unambiguous (red box closed)**	actual (fox_Q)	2.86%(7.55%)
		Impossible1 (chick_Q)	2.86%(7.55%)
		Impossible2(turtle_Q)	0.95%(3.93%)
	True Belief-ambiguous (green box closed)	actual (turtle_Q)	4.76%(8.64%)
		Impossible1(chick_Q)	2.38%(5.92%)
		Impossible2(fox_Q)	3.33%(7.88%)
	Underspecified Belief (yellow box closed)	actual(chick_Q)	4.76%(9.54%)
		possible(turtle_Q)	3.81%(9.12%)
		Impossible(fox_Q)	3.33%(10.54%)
		Total	3.23%(1.19%)
SELF	**True Belief-unambiguous (red box closed)**	actual (fox_Q)	1.43%(4.73%)
		Impossible1 (chick_Q)	5.24%(8.83%)
		Impossible2(turtle_Q)	1.90%(7.85%)
	True Belief-ambiguous (green box closed)	actual (turtle_Q)	4.29%(10.95%)
		Impossible1(chick_Q)	2.86%(7.55%)
		Impossible2(fox_Q)	3.81%(9.97%)
	Underspecified Belief (yellow box closed)	actual(chick_Q)	3.33%(7.88%)
		possible(turtle_Q)	3.81%(8.17%)
		Impossible(fox_Q)	5.71%(13.37%)
		Total	3.60%(1.41%)

S4.2.2 RT data exclusion rates

Table S4.3. *The mean percentage of trials (SD) excluded from the RT analyses in Experiment 1, 2 and 3 (excluding those participants whose mean reaction time was 2SD above or below the average, calculated for all trials).*

	belief	alternative type	Experiment 1 N=(33)	Experiment 2 (N=31)	Experiment 3 (N=33)
OTHER	True Belief	actual	3.54% (6.92%)	3.23% (7.96%)	2.08% (5.60%)
		possible	4.55%(9.57%)	2.15% (5.68%)	4.17% (8.47%)
		impossible	2.02%(5.52%)	3.23% (7.96%)	3.13% (6.61%)
	Underspecified Belief	actual	3.54%(6.92%)	3.23% (6.69%)	3.65% (7.00%)
		possible	2.53%(9.43%)	2.15% (7.13%)	4.17% (8.47%)
		impossible	5.56%(10.76%)	1.61% (5.01%)	2.08% (5.60%)
		Total	3.62%(1.29%)	2.60%(0.71%)	3.21%(0.96%)
SELF	True Belief	actual	2.02%(5.52%)	3.76% (7.08%)	4.17% (8.47%)
		possible	4.55%(8.61%)	1.61% (5.01%)	2.60% (6.15%)
		impossible	5.56%(7.98%)	0.54% (2.99%)	4.17% (9.47%)
	Underspecified Belief	actual	1.01%(4.04%)	1.61% (5.01%)	5.73% (9.09%)
		possible	4.55%(8.61%)	2.15% (7.13%)	3.13% (6.61%)
		impossible	5.05%(8.82%)	2.15% (5.68%)	2.60% (6.15%)
		Total	3.79%(1.82%)	1.97%(1.06%)	3.73%(1.21%)

Note: In Experiment 2 and 3 no trials had to be excluded due to the RT data being above or below RT mean +/– 2SD. Experiment 1: percentages are sometimes lower than in Table S1 due to the different sample size.

Table S4.4. *The mean percentage of trials (SD) excluded from the RT analyses in Experiment 4 (excluding those participants whose mean reaction time was 2SD above or below the average, calculated for all trials).*

	belief	alternative type	Experiment 4 (N=33)
OTHER	True Belief-unambiguous (red box closed)	actual (fox_Q)	3.54%(8.08%)
		Impossible1 (chick_Q)	3.03%(7.74%)
		Impossible2(turtle_Q)	0.51%(2.90%)
	True Belief-ambiguous (green box closed)	actual (turtle_Q)	4.55%(8.61%)
		impossible(chick_Q)	2.02%(5.52%)
		Impossible(fox_Q)	4.04%(8.36%)
	Underspecified Belief (yellow box closed)	actual(chick_Q)	4.04%(8.36%)
		possible(turtle_Q)	3.03%(7.74%)
		Impossible(fox_Q)	1.52%(6.41%)
		Total	2.92%(1.33%)
SELF	True Belief-unambiguous (red box closed)	actual (fox_Q)	2.02%(5.52%)
		Impossible1 (chick_Q)	5.05%(8.82%)
		Impossible2(turtle_Q)	1.01%(5.80%)
	True Belief-ambiguous (green box closed)	actual (turtle_Q)	3.54%(8.08%)
		impossible(chick_Q)	3.03%(7.74%)
		Impossible(fox_Q)	2.53%(6.07%)
	Underspecified Belief (yellow box closed)	actual(chick_Q)	2.02%(5.52%)
		possible(turtle_Q)	3.03%(7.74%)
		Impossible(fox_Q)	4.55%(8.61%)
		Total	2.97%(1.27%)