Cognition 249 (2024) 105812

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full Length Article

Flexible social monitoring as revealed by eye movements: Spontaneous mental state updating triggered by others' unexpected actions

Dóra Fogd *, Natalie Sebanz, Ágnes Melinda Kovács

Department of Cognitive Science, Central European University, Vienna, Austria

ARTICLE INFO

Keywords:
Theory of mind
Spontaneous mentalizing
Mental state update
Transfer of the updated content
Anticipatory looking

ABSTRACT

Successful interactions require not only representing others' mental states but also flexibly updating them, whenever one's original inferences may no longer hold. Such situations arise, for instance, when a partner's behavior is incongruent with one's expectations. Although these situations are rather common, the question whether people update others' mental states spontaneously upon encountering unexpected behaviors and whether they use the updated mental states in novel contexts, has been largely unexplored. We addressed these issues in two experiments. In each experiment participants first performed an anticipatory looking task, reacting to a virtual 'partner', who categorized pictures based on their ambiguous or non-ambiguous color. Importantly, to perform the task participants did not have to track their partner's perspective. Following a correct categorization phase, the 'partner' started to systematically miscategorize one of the ambiguous colors (e.g., as if she would now believe that the greenish blue is green). We measured how participants' anticipatory looking preceding the partner's categorization changed across trials. Afterward, we asked whether participants implicitly transferred their knowledge about the partner's updated perspective to a new task. Finally, they performed an explicit perspective-taking task, to test whether they selectively updated the partner's perspective, but not their own. Results revealed that correct anticipations started to emerge only after a few miscategorizations, indicating the spontaneous updating of the other's perspective regarding the miscategorized color. Signatures of updating emerged somewhat earlier when the partner made similarity judgments (Experiment 2), highlighting the subjective nature of her decisions, compared to when following an explicit color-categorization rule (Experiment 1). In the explicit perspective-taking task of both experiments, roughly half of the participants could categorize items according to the partner's (spontaneously updated) perspective and also used their partner's updated perspective in the implicit transfer task to some degree, while they were the ones who displayed more pronounced anticipatory patterns as well. Such data provides strong evidence that the observed changes in anticipatory looking reflect spontaneous and flexible mental state updating. In addition, the findings also point to a high individual variability both in the updating of attributed mental states and the use of the updated mental state content.

1. Introduction

From catching a ball in a basketball game to taking part in an everyday conversation, efficient social interactions require making correct predictions about what others will do or say. Most often humans perform such predictions by taking into account the unobservable mental states of others, expecting them to act in line with what they want, see, believe or know. Importantly, the capacity to represent other agents' mental states, usually termed theory of mind (ToM) or mentalizing, entails understanding that others' points of view may not coincide with the true state of affairs and may differ from one's own (Premack & Woodruff, 1978). If, for example, in a basketball game, a player believes that her co-player did not see the opponent on the right, she will take this into account when formulating expectations regarding the other's future move and will adjust her own movements accordingly. In a similar vein, if players observe their team leader deviating from the agreed-upon strategy, they will assume that the leader has a good reason to do so, even if the exact reason might not be that straightforward when 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, as well as endorsing a discrepancy between her mental state and their own (e.g., 'maybe she has a new strategy in

* Corresponding author at: Department of Cognitive Science, Central European University Private University, Quellenstasse 51, 5th floor, 1100 Wien, Austria.
E-mail address: fogdd@ceu.edu (D. Fogd).

https://doi.org/10.1016/j.cognition.2024.105812
Received 9 August 2023; Received in revised form 9 May 2024; Accepted 10 May 2024
Available online 18 May 2024

mind and aims to attack taking advantage of the situation that ensued'). How humans make such rapid and seemingly complex inferences, i.e., revise their own beliefs about others' beliefs, and, consequently, their expectations regarding others' actions upon encountering an unexpected behavior has been so far largely unexplored. Such a flexible monitoring of our social partners, in particular, tracking their mental states and revising them if needed, is of vital importance as this allows humans to adapt and thrive in the continuously changing social milieu.

To date, most of the research investigating ToM abilities has focused on the question whether adults and children can take into account others' different beliefs or knowledge states, with very little attention devoted to flexible belief revision. While in the traditional ToM tasks participants are explicitly asked to predict others' behavior on the basis of their mental states (Baron-Cohen, Leslie, & Frith, 1985; Birch & Bloom, 2007; Dziobek et al., 2006), more recently, attention has shifted to nonverbal or 'implicit' measures of ToM abilities, such as reaction times or anticipatory looking. Studies from the last 15 years suggest that human adults may compute the content of others' mental states spontaneously. Their responses are affected by what another agent sees and believes in tasks that do not require tracking the mental states of the other (Buttelmann & Buttelmann, 2017; Elekes, Varga, & Király, 2016; Kovács, Téglás, & Endress, 2010; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; Surtees, Apperly, & Samson, 2016; van der Wel, Sebanz, & Knoblich, 2014). They also display anticipatory eye movements indicative of attributing a false belief to another person, even when they are not instructed to predict the other agent's actions (Schneider, Bayliss, Becker, & Dux, 2012; Schneider, Nott, & Dux, 2014).

These results and the interpretation that they reflect genuine mentalizing abilities have generated considerable debate in the past decade. For instance, it was argued that the implicit tracking of mental states may not be part of the fully-fledged ToM system, but rather it may be subserved by a low-level system that relies on different kinds of representations (Butterfill & Apperly, 2013). Other accounts have suggested that the effects observed in the implicit mentalizing tasks may merely reflect the operation of low-level attentional processes (Heyes, 2014; Phillips et al., 2015). Failures to replicate some of the early findings of studies targeting spontaneous ToM, in particular, those relying on anticipatory looking (Burnside, Ruel, Azar, & Poulin-Dubois, 2018; Kampis, Karman, Csibra, Southgate, & Hernik, 2021; Kulke, Reiß, Krist, & Rakoczy, 2018), have also led many authors to question the original, mentalistic interpretation of these results.

Although the debate is far from being settled, there is a growing amount of evidence suggesting that perspective-taking and mental state attribution may take place spontaneously in humans (Freundlieb, Kovács, & Sebanz, 2018; Seow & Fleming, 2019; Ward, Ganis, & Bach, 2019). Earlier findings seem to replicate under conditions that rule out the low-level processes proposed by Heyes (2014) (Furlanetto, Becchio, Samson, & Apperly, 2016; though for a nonreplication see: Conway, Lee, Ojaghi, Catmur, & Bird, 2017) or alternative explanations proposed, for instance, by Phillips et al. (2015) (see El Kaddouri, Bardi, De Bremaeker, Brass, & Wiersema, 2020), thus providing support for mentalistic accounts. Neuroimaging studies also suggest that implicit mentalizing shares neural substrates with (Hyde, Aparicio Betancourt, & Simon, 2015; Naughtin et al., 2017) and relies on the same computations as explicit ToM (Nijhof, Brass, Bardi, & Wiersema, 2016, but see: Schneider, Slaughter, Becker, & Dux, 2014). With respect to the above-mentioned nonreplications of the anticipatory looking findings (Kampis et al., 2021; Kulke et al., 2018), other results suggest that these failures may reflect limitations of the specific paradigms used in these studies, rather than limitations of the spontaneous ToM capacity of human adults. In particular, eyetracking studies using the visual world paradigm consistently report an early sensitivity of adults to the false beliefs and discrepant visual perspectives of other agents (Cane, Ferguson, & Apperly, 2017; Hanna, Tanenhaus, & Trueswell, 2003; Symeonidou, Dumontheil, Ferguson, & Breheny, 2020). Taken together these findings suggest that human adults may readily engage in mentalizing in various social contexts.

Importantly, however, 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 might no longer hold – they have become outdated or might have been completely wrong. Such situations arise, for example, when one observes another agent witnessing an event that should provoke a change in her knowledge state, or when one acquires a new piece of information that is incompatible with an earlier attributed mental state. Notably, these two situations differ markedly in how updating is performed. In the first case, one has to perform the updating 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, Onishi, Baillargeon, & Fisher, 2008; Tauzin & Gergely, 2019), suggesting that such computations might take place spontaneously in humans. Much less is known about how people update other agents' mental states when the situation requires retrospective recomputation of the previously attributed belief content. Recent studies suggest that children can correct their earlier attributions (from true to false belief, or from false to true), spontaneously, from very early on, when provided with new information that sheds new light on what events the agent could have witnessed before (Király, Oláh, Csibra, & Kovács, 2018; Király, Oláh, & Kovács, 2023).

Crucially, situations that necessitate the retrospective update of another agent's mental state arise not only when one receives a new piece of information about the circumstances in which the other's current beliefs were formed. 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, more specifically from one's own failure to correctly predict the other's actions (e.g., the team leader's next move). Although such situations are quite frequent 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. Considering humans' remarkable capacity to infer the goals, beliefs and stable preferences of any agentive entity, merely based on the actions it performs and the environmental constraints of those actions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009), they should also be able to efficiently update previously attributed beliefs on the basis of such information.

Everyday intuition suggests that people readily update other agents' mental states upon observing actions that violate their expectations, without the need for being asked to do so. They quickly intervene if someone behaves as if being unaware of an imminent danger, 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 behavior. It is not entirely clear though whether adults perform such computations spontaneously, whenever they encounter an unexpected action, or only when it is absolutely necessary, to avoid harm or the breakdown of an ongoing interaction. A recent study, targeting a different question, specifically how to disentangle the effects of selective attention and ToM use in a referential communication task, suggests that adults may spontaneously infer that another person has more information than they originally thought, from the way the person behaves in a situation (Rubio-Fernández, 2017). In this study, most of the adult participants noticed that, on some trials, the director's verbal descriptions of the target objects (e.g., 'the orange fish') did not match his/her alleged knowledge state (s/he was supposed to see one fish), and, by the end of the task, started to look more towards those allegedly hidden objects (e.g., the other fish) that were actually seen by the director. This can be taken as indirect evidence that, at least in interactive contexts, adults may spontaneously update the content of another agent's mental state, when his/her actions violate their expectations. Nevertheless, as the study was designed to investigate a different research question, the interpretation of the observed change in participants' looking behavior is far from straightforward. In particular, it is unclear whether the increased looking at the allegedly hidden objects is due to mentioning a color contrast and the disambiguation process, or due to updating the director's perspective. Hence, these findings do not allow firm conclusions to be drawn regarding how mental state updating may take place upon observing an unexpected behavior: what are the necessary preconditions for such a recomputation, how much evidence is needed to trigger the process, and most importantly, to what extent people flexibly revise their expectations regarding the others' behavior upon performing such computations.

Thus, the present study aimed to address these questions. In specific, we aimed at directly investigating whether people update the content of another agent's mental state and revise their expectations about her future behavior spontaneously, even if they could solve their task without doing so, upon observing her repeatedly acting in a way that does not follow from their initial assumptions regarding what she sees or knows. In addition, we also wished to explore how such a process may unfold, in specific, whether updating takes place already after observing a few unexpected actions.

To this end, we developed an anticipatory looking task in which participants were asked to react to another person, 'the partner', who was performing a color-based categorization task in another room, involving blue and green colors that were easy to categorize (fully unambiguous blue or green) or somewhat harder to assign to a color category (slightly ambiguous: blueish or greenish, see Fig. 1a), while their eye movements were recorded. At a certain point in the task the 'partner' started to categorize one of the intermediate, hence somewhat ambiguous colors incorrectly, assigning it to the opposite category the particular color belonged to (i.e. 'greenish' as blue or 'blueish' as green). We measured whether participants' anticipatory looking behavior changed in response to these actions, in particular, when they started to correctly anticipate the 'partner's' future behavior on those trials where the other acted in a rule-incongruent manner (e.g., put the somewhat green blue - 'blueish' - items in the green box), signaling the updating of the other's perspective regarding that specific color. Unbeknownst to participants the 'partner' was a computer program. In each trial the 'partner' first selected one out of two pictures, depicting animals, following the auditory instructions of an unseen 'director', then categorized the picture based on the color of its frame (blue or green), by clicking on the box with the respective color label. This clicking resulted in the lighting up of the selected box and marked the beginning of the participants' turn. Participants' task was to simply click on the box that lit up, as fast as possible. They were not instructed to track the 'partner's' mental states or anticipate her behavior. Nevertheless, they could do so by spontaneously looking at the respective box before it lights up. Our crucial experimental manipulation was that after a period of correct categorization, the 'partner' started to systematically miscategorize one of the two ambiguous colors (i.e. incorrectly categorize the greenish as blue and the blueish as green), as if she had changed her mind regarding which category that specific color belongs to. We use the term miscategorization as these actions differed from how the majority of people would categorize the particular color (based on which of the two colors dominate the compound).

To measure whether and how participants update the other's mental state (how the partner represents a particular color) upon observing the miscategorizations and revise their future expectations regarding the other's actions, we recorded anticipatory looking towards the boxes in the period preceding the other's decisions (signaled by the lighting up of the box) and analyzed how looking behavior changed following the change in the other's behavior. We hypothesized that, if participants spontaneously update the other's mental state, then, after the observation of some unexpected actions ('miscategorizations'), they will start to look more towards the box that is an incorrect choice from a first-person perspective, but corresponds to the 'partner's' belief regarding the frame's color (i.e. correct choice from his/her point of view), prior the other would actually select this box, as compared to how much they look towards the box that is incorrect both from first- and third-person-perspective in this period in the control condition, when the frame has the other ambiguous color that is categorized properly by the 'partner'.

However, one may argue that the expected changes in anticipations may also reflect the revision of a rule participants had acquired (a specific color associated with a specific box becomes later associated with the other box) and not just the updating of the 'partner's' mental state (i.

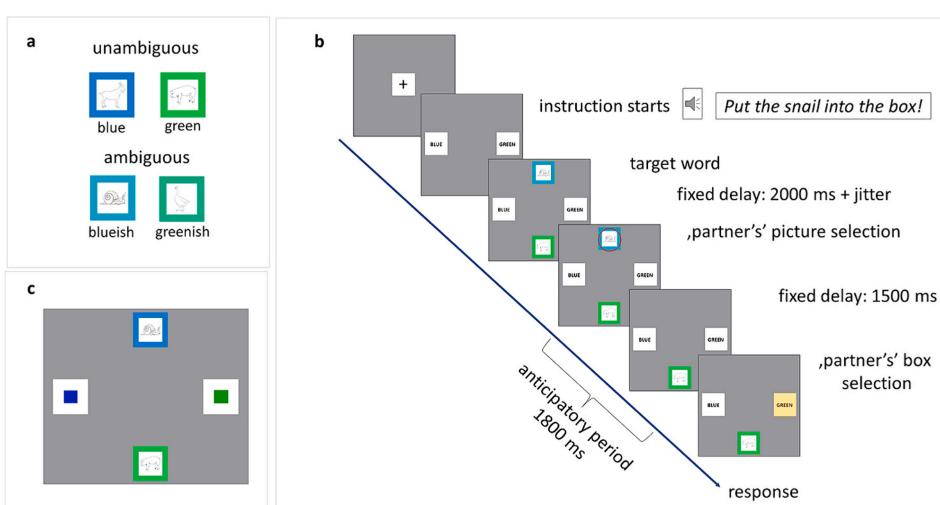


Fig. 1. Schematic illustration of the anticipatory looking task. a) Stimuli and (b) trial structure of the anticipatory looking task used in Experiment 1 and Experiment 2. The 'partner' first selected the animal named in the instruction and then categorized the picture according to the frame's color, 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. (c) In Experiment 2, instead of the color labels, blue and green colored squares were displayed on the boxes and the 'partner' had to put the picture into the box which was judged more similar in color to that of the selected picture's frame.

e., 'she thinks this greenish color is blue'). To exclude this possibility, 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 novel geometric shapes, having the same color as the picture frames, based on their color, either from their own or, in separate blocks, from the 'partner's' perspective. We conjectured that if participants can take into account how the other perceives the miscategorized color when categorizing items of that specific color from the 'partner's' perspective, while keeping the original categorization rule from self-perspective, this can be taken as clear evidence that prior changes in their looking behavior reflect the updating of the other's mental state and not that of a nonmentalistic rule.

In addition to the anticipatory looking and the explicit perspective-taking task, participants also performed a third task (administered second during the study) to investigate whether they would use their newly acquired knowledge about the other agent (i.e., her updated 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 specific color), one may expect people to generalize what they have learnt about the other in one context to future social interactions.

To investigate whether this is the case, following the anticipatory looking task, participants performed a virtual coordination task in which they could use their newly acquired knowledge about their partner (how the other sees one of the ambiguous colors), to predict where the other expects them to pass an object (i.e., at the blue or the green 'gate'). In this 'implicit transfer task', participants could optimize their joint performance with the partner (to move an object to a goal location as fast as possible) by taking into account what they have learned about how the other sees one of the ambiguous colors in the anticipatory looking task. Our critical question was whether participants would pass the 'previously miscategorized colored' figures at the gate that is correct from the other's perspective, but not from their own. If so, it would provide evidence for a spontaneous ('implicit') transfer of the updated belief content to another context and, hence, a further proof for the mentalistic account of the predicted change in participants looking behavior in the anticipatory task.

To sum up, in Experiment 1 participants performed three tasks, in the following order: an anticipatory looking task, an implicit transfer task, and an explicit perspective-taking task. To test whether updating of another agent's mental state is easier if the revision does not require the overcoming of an explicit categorization rule, we also ran a second experiment. Experiment 2 differed from Experiment 1 only in one aspect: instead of categorizing the pictures based on their frame's color, the 'partner' had to make similarity judgments, to highlight more the subjective nature of her decisions and thereby possibly facilitate the process of mental state updating.

2. Experiment 1

Experiment 1 investigated whether adults spontaneously update the content of another agent's mental state and revise their expectations about her future behavior. The study comprised three tasks participants performed in a fixed order: an anticipatory looking task, an implicit transfer task, and an explicit perspective-taking task. The first and the third tasks served to test our hypothesis regarding the updating of others' mental states on the basis of their behavior and the second investigated whether people generalize their newly acquired knowledge about the other's updated mental state to subsequent interactions.

2.1. Methods

2.1.1. Participants

37 university students were recruited for the experiment via a student job agency and the university's research participation system (SONA systems). The target sample size was determined based on the only anticipatory looking study we could find that used a multiple-trial design to investigate implicit ToM in adults (Schneider, Bayliss, et al., 2012). The data of 7 participants was excluded 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, in which the 'partner' first categorized the ambiguous color correctly[1]: n=1). Thus, the final sample consisted of 30 participants (age: M= 23.77, SD=4.34, 16 males). All of them were right-handed and had normal or corrected-to-normal vision (no reported color blindness). The study was approved by the EPKEB United Ethical Committee for Research in Psychology, Hungary; participants signed informed consent and received monetary compensation (equivalent to 7 euros) for their participation.

2.1.2. Anticipatory looking task

2.1.2.1. Stimuli. Visual stimuli consisted of two 'boxes' – two white squares of the same size (subtending 4.27° of visual angle horizontally and vertically), with the Hungarian words for 'BLUE' and 'GREEN' displayed in their middle -, and line drawings of four animals, presented on a white background (having the same size as the boxes), with a colored frame (see Fig. 1a for examples). Stimuli were displayed on a plain grey background.

The colors 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. Importantly, one shade of each color was slightly harder to categorize (as blue or green) than the other by virtue of being somewhat closer to the other color in the RGB color space (i.e. being an intermediate color). The aim of this manipulation was to make it credible for participants that another person may perceive these colors in a different way. We refer to these harder-to-categorize colors as ambiguous colors and the other two colors as unambiguous colors in the paper. Independent rating data confirmed that while the ambiguous colors could be clearly categorized as blue or green, based on their dominant color (by >90% of participants), participants' confidence was lower for these than for the unambiguous colors (for details see the Supplementary Materials, Section A).

The auditory stimuli comprised four sentences that were identical except the last word, which identified one of the four animals (e.g., '[Put] [into the box] [*the target animal*]'[2]). Target words (goat, pig, snail, goose, in Hungarian; 680 ms long each) were bisyllabic and started with a consonant.

2.1.2.2. Apparatus. Eye movements were recorded by a Tobii X60 eyetracker (Tobii Technology, Sweden). We used a chinrest with a forehead support (SR Research Head Support), ensuring a 52 cm viewing distance. 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.

2.1.2.3. Procedure. Upon arrival, participants were told that they would perform the experiment together with another person, but in separate rooms, connected via Internet, and they were allocated the role

[1] Correct anticipation was defined as a mean proportion of looking > 0.50 on both kinds of ambiguous trials, either in the 1800 ms time window before box selection or in the 2000 ms time window prior picture selection (for details see the Supplementary Materials).

[2] This is the possible word order in Hungarian.

of 'player B' (in a seemingly random manner). They were told that, since their partner is late, unlike usual, they will not receive the instructions together. The anticipatory task was introduced as a speeded 'action coordination' task, in which the 'partner' had to perform the first step ('react' to a linguistic stimulus) and the participant the second step ('react' to the action of another person). The 'partner's task was to put the picture of the animal named by an unseen 'director' into the box matching the picture's frame color (explicitly telling them that '*if it is blue then they have to click on the BLUE box, if it is green then on the GREEN box*'). More specifically, the 'partner' first had to select the animal, by clicking on it, which led to the appearance of a red circle around the picture, then categorize it according to the picture's frame color, by clicking on one of the two presented boxes, which resulted in the lighting up of the selected box. The participant's task was to react to the other as fast as possible, by clicking on the box that lit up. Since participants always had to act on the box that was selected by their 'partner', we label these as 'target' locations and the other box as 'nontarget' location, throughout the paper.

To ensure that participants believed that they perform the task together with another human, the instructions (describing the task of both 'players') 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 this was 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 (repeated if necessary). Then, the instruction was repeated, and the task started.

The task consisted of a familiarization and a test phase. The trials had the same structure in both phases (see Fig. 1b). Each trial started with a 1500 ms central fixation cross on a white rectangular background (subtending 4.40° of visual angle in height and width). Then the two boxes were presented, on the left and right, along with the verbal instruction for the 'partner' (*'Put in the box* [the target animal]*;* 1282 ms). Upon displaying the target word, two pictures depicting an animal (the target and a distractor) appeared on the upper and lower parts of the screen, at equal distance from the center. One of the pictures always had a blue or a blueish frame, the other a green or a greenish one. The 'partner's 'target selection period' lasted for 2000-2750 ms, with the jitter (randomly varying between 0-750 ms, in 250 ms bins) introduced to make it more credible that the decision is made by another human. Afterward, a red circle appeared around the target picture (300 ms), indicating the partner's choice. Then the target picture disappeared, and the screen remained still for 1500 ms. The 'partner' selection of the box was indicated by lighting up the respective box (300 ms). The onset of this event marked the beginning of the participant's response period, which lasted for a maximum of 2800 ms or until response was given. Trials were separated by 2000 ms intertrial intervals.

The familiarization phase consisted of 16 trials. Depending on the color of the target picture's frame the trials could belong to one of four conditions - unambiguous blue, unambiguous green, ambiguous blue and ambiguous green – each presented 4 times, in the same pseudorandomized order for each participant (for details see the test phase). Importantly, during this phase, the 'partner' always acted as expected, i.e., selected the appropriate box. As a result, during these trials, participants could develop clear expectations regarding how the events unfold and, consequently, develop anticipatory looking towards the appropriate locations.

After the familiarization, 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 (but not the other three) colors. Such a manipulation yielded the following four experimental conditions: 'unambiguous blue', 'unambiguous green', 'properly categorized ambiguous' and 'miscategorized ambiguous'. The miscategorized ambiguous color (blueish or greenish) was kept constant throughout the test phase and was counterbalanced across participants.

Participants received four blocks of 32 test trials, separated by short, self-paced breaks, with each block containing eight trials from each condition, resulting in a total of 32 trials per condition. The order of the trials within the blocks was pseudorandomized. For further details see the Supplementary Materials.

2.1.3. Implicit transfer task

2.1.3.1. Stimuli and apparatus. Stimuli consisted of two white squares of the same size (the 'gates', subtending 4.27° of visual angle in height and width), with grey bars at their lateral sides, and the Hungarian words for 'BLUE' and 'GREEN' written on them, that were displayed in the right and the left corners of the bottom of the screen, another, bigger white square (the 'box', having the same size as the 'gates') with the letter D on it (denoting the Hungarian word for 'box'- 'Doboz'), presented in the middle on the top of the screen, and four colored geometric shapes (labeled as 'figures', each subtending 3.30° of visual angle in diameter). The colors were identical to the ones used in the anticipatory looking task. 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.

2.1.3.2. Procedure. 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 (performing 'complementary' actions), 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 'waited' for the figure. The 'partner's' task was to finish the round by moving the figure into the 'box' as fast as she could. Participants were told that the two of them would 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 a central fixation cross, presented for 1000 ms. Then the two 'gates' and the 'box' appeared (for the exact layout see Fig. 2a). 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 (at equal distance from the 'gates'). After the 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. Trials were separated by 1000 ms intertrial intervals. 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 color of the figure, trials could belong to one of the following four experimental conditions: unambiguous blue, unambiguous green, previously miscategorized and previously properly categorized ambiguous trials.

The task started with eight introductory trials (two from each condition), which was followed by a message on the screen warning participants that they and their 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 the consideration of the other's perspective to improve their performance. Following this, participants received two blocks of 32 test trials, with each block containing eight trials from each

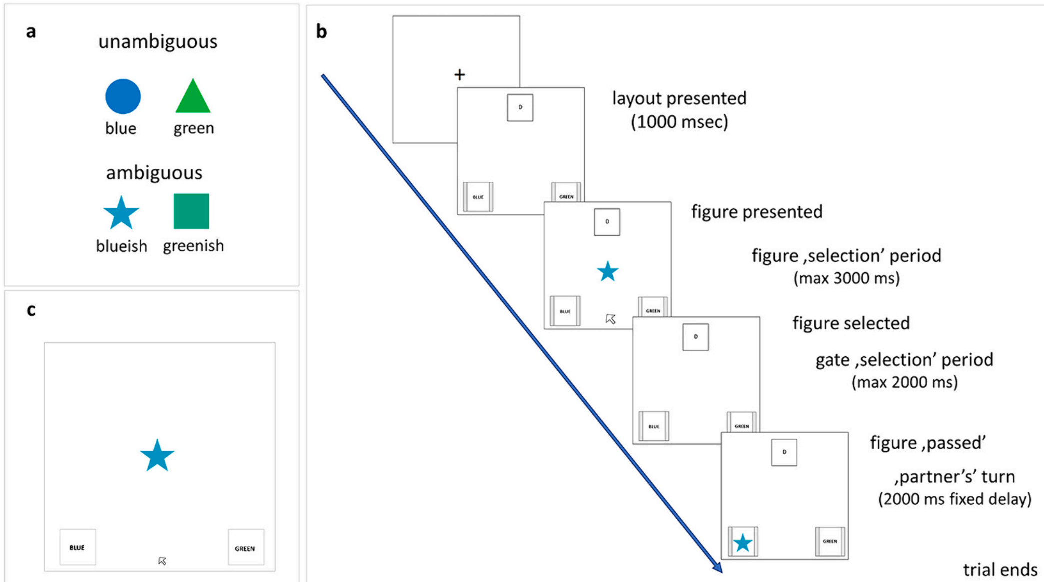


Fig. 2. a) Stimuli and (b) trial structure of the implicit transfer task. The participant's task was to first click on the geometric shape and then on the gate where the 'partner' waits for his/her pass. The 'partner' then had to move the item into the box as fast as possible. Participants did not see their 'partner''s' actions. (c) Layout of the explicit perspective taking task. Participants had to categorize the geometric shapes (figures) of the implicit transfer task and the animal pictures of the anticipatory looking task either from their own (self trials) or from their 'partner''s' (other trials) perspective (in separate blocks). Their task was to first click on the presented item then on the appropriate box. In Experiment 2 blue and green colored squares were displayed on the boxes instead of the color labels, both in the implicit transfer and in the explicit perspective taking task.

condition. (For further details see the Supplementary Materials).

2.1.4. Explicit perspective-taking task

2.1.4.1. Stimuli and apparatus. Stimuli consisted of the animal pictures and the two 'boxes' used in the anticipatory looking task and the geometric figures, used in the implicit transfer task. The apparatus and the settings were the same as before.

The primary focus of interest was whether participants will be able to correctly categorize the geometric figures used in the implicit transfer task from the other's perspective (e.g., whether they would categorize the somewhat ambiguous green-colored figures as blue, if the 'partner' categorized the animal pictures with this frame color as blue in the anticipatory looking task). Importantly, participants never saw how the other categorized the geometric figures in the previous tasks, hence appropriate categorization was not possible simply by relying on their memory of how their 'partner' acted earlier in the experiment.

2.1.4.2. Procedure. Participants were presented with four blocks of 32 trials, two of which required them to categorize the stimuli according to their color, from their own perspective, and the other two to perform the categorization from their partner's perspective ('as their partner would categorize' the given item).

Trials started with a central fixation cross (presented for 1000 ms), Then the two 'boxes' of the anticipatory task appeared in the bottom left and right corner of the screen. 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'). Following the response, the geometric figure/animal picture disappeared, and participants had a maximum of 2000 ms to perform the categorization, by clicking on one of the boxes. After the selection of the box, the geometric figure/animal picture appeared inside (300 ms), then the trial ended. Trials were separated by a 1000 ms intertrial interval. Stimulus presentation was blocked, with the geometric figures being presented first and the animal pictures second within each pair of blocks, to avoid that on the other perspective block participants' categorization of the geometric figures is influenced by their memories of how the 'partner' categorized the colored frames of the animal pictures. Trials could belong to one of four conditions (unambiguous blue, unambiguous green, previously miscategorized and properly categorized ambiguous color). Each block contained eight trials from each condition, with items counterbalanced within the conditions. For further details see the Supplementary Materials.

Following the explicit perspective-taking task, we asked verbal feedback from the participants. In specific, we asked 1) whether they have noticed anything peculiar in their partner's behavior (and, if yes, what is their interpretation for it) and 2) to what extent did they believe that they were interacting with another human. Given that these questions themselves could elicit a revision of participants' original beliefs (e.g. raise doubts that they were playing with an actual human), the verbal responses participants provided at this point were not used for data exclusion or creating subgroups for the data analyses. After filling out the questionnaire, participants were extensively debriefed, revealing that they were playing with a computer.

The whole experiment lasted for about 80 minutes.

2.1.5. Data analysis

The data and the code used for the analyses are available in the Open Science Framework (OSF) repository (https://osf.io/wka2s/).

2.1.5.1. Anticipatory looking task. Eyetracking analyses were based on the raw gaze data of the participants. Anticipatory looking was analyzed in the 1800 ms long period after the onset of the 'partner's' picture selection until the lighting up of the target box. 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 nontarget box. Importantly, on the 'miscategorized' test trials the 'target AOI' was the box that did not correspond to the task rules but corresponded to how the 'partner' perceived the particular color.

Our main analyses focused on the *proportion of looking* at the target box in this time window of the test trials. Proportion of looking 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 ('no-anticipation' trials). In addition, we also analyzed at which box the participant *look*ed *first* in this time period, to check whether results obtained with a different measure corroborate our findings in the main analyses. First look at the target box was coded as 1, at the nontarget box as 0 (for the detailed results see the Supplementary Materials Section C).

Data from the familiarization phase was analyzed by averaging the proportion of looking separately for the ambiguous and unambiguous trials (combining the blue and green trials), to test whether participants have formed initial expectations about the other's actions in this phase, regarding both the ambiguous and unambiguous colors. Participants for whom the proportion of anticipatory looking in familiarization was lower than 0.5 on the ambiguous trials were excluded from the further analyses. The results of this phase are presented in the Supplementary Materials.

Statistical analyses were conducted using R (Version 4.3.1; R Core Team, 2023) and RStudio (Version 4.3.1, Posit Team, 2023). Test phase anticipatory looking was analyzed with generalized linear mixed-effects models, to maximize the amount of data we could include in our analyses, using the *glmmTMB* (*glmmTMB* package, Brooks et al., 2017) and the *glmer* functions (*lme4* package, Bates, Mächler, Bolker, & Walker, 2015) of R, for the analysis of the proportion of looking and the first look data, respectively. In all analyses participants were treated as random and condition and block as fixed effects (setting the properly categorized ambiguous color condition and block 1 as reference). After running the full models ($DV \sim condition * block + (1|id)$), we tested for the effect of condition separately in each block ($DV \sim condition + (1|id)$), to investigate at which timepoint conditions start to differ. Given that proportions are best modeled with beta distribution, due to the bounded nature of the data (see e.g. Ferrari & Cribari-Neto, 2004), and the fact that our proportion of looking data contained large number of zeros and ones (see Fig. S4), that, however, could not be considered as qualitatively different from the rest of the values (i.e. data generated by a different process or source), we decided to analyze the data specifying *beta_family* as distribution with logit link function[3], following the transformation of our proportion of looking variable. Transformation was necessary to fit the data in the (0,1) interval and was performed applying the formula suggested by Smithson and Verkuilen (2006); $PL\_recod = (PL*(N - 1) + .5)/N)$). First look data was simply treated as binomial, setting logit link function, using the *bobyaga* optimizer (with boundary tolerance set at .001) to ensure convergence. Model fitting was tested with k-fold cross-validation, performed using the *cv* package of R (Fox & Monette, 2024), using k=3 clusters. For the ease of reading, we report only the results of the ANOVA table for the full models (obtained using the *car* package, Fox & Weisberg, 2019) in the main text and present the parameter estimates and the cross-validation criteria in the respective sections of the Supplementary Materials.

Importantly, participants' own perspective and the task-rule (how the pictures *should* be categorized), which remained valid throughout the task could be expected to exert a strong influence on anticipatory looking, similar to the 'pull of the real', demonstrated in other false-belief anticipatory looking paradigms (see e.g., Schneider, Slaughter, et al., 2014). Therefore we did not expect a full switch in the anticipatory looking, from the rule-congruent location to the location that corresponded to the other's perspective but was incongruent with the task-rule (and the first-person perspective), on the miscategorized trials. What we predicted was rather an increase in the anticipatory looks towards the location consistent with the 'partner's' (updated) perspective, i.e. more looking towards this location than what could result from the scanning of the rule-incongruent (nontarget) location in a slightly ambiguous situation. Based on these considerations, we decided, a priori, to compare the proportion of looking to the target box on the miscategorized ambiguous trials to the proportion of looking to the nontarget box on the other, properly categorized ambiguous color trials. Since our hypotheses concerned the difference between the two ambiguous color conditions, all the aforementioned analyses were run without including the other two conditions. (For the proportion of looking results in the two other – unambiguous - conditions see the Supplementary Materials Section E.)

2.1.6. Implicit transfer and the explicit perspective-taking task

2.1.6.1. Accuracy analyses. For both tasks, the primary dependent measure was accuracy on the previously miscategorized ambiguous trials, with comparisons made to accuracy on the previously properly categorized ambiguous trials. Importantly, passing/categorizing the presented item according to its actual color, i.e. congruent with one's own perspective was always coded as 1, independent of condition, as this could be considered the correct choice on most of the trials. Therefore, taking into account the other's (updated) perspective, on the previously miscategorized ambiguous trials, is reflected in lower (and not higher) scores in both tasks.

In case of the explicit perspective taking task, accuracy was analyzed with generalized linear mixed-effects models, specifying binomial distribution (with logit link function). Just like in the eyetracking analyses, participants were treated as random and perspective and condition as fixed effects (with the self-perspective and the previously properly categorized ambiguous color condition set as reference). Models were run separately for the animal pictures and the geometric figures, as these items differed in whether or not participants saw how their 'partner' categorized them in the anticipatory looking task, hence in the cognitive processes participants' categorization could possibly reflect. After running the full models ($DV \sim perspective * condition + (1|id)$), as a follow-up tests, we fitted models separately for the two perspectives. To ensure convergence we used the *bobyga* optimizer with boundary tolerance set at .001 in all of the analyses.

Since testing our hypothesis regarding performance in the implicit transfer task required a single comparison, in this task, accuracy was analyzed by simply averaging the responses across the 16 trials (with introductory trials not included in the analyses) and then comparing the hit rates of the two conditions. Due to the nonnormal distribution of the data analyses were run using Wilcoxon Signed Rank Tests, using the *rstatix* package of R (Kassambara, 2023; calculating the Z-score using the qnorm(p/2) formula).

2.1.6.2. Latency analyses. To gain a more fine-grained picture of the processes underlying the participants' decisions, we also analyzed picture selection latencies (defined as the time it took for participants to click on the presented item from the appearance of the cursor)[4], with linear mixed-models in case of the explicit perspective taking task (specifying the same random and fixed effects as above), using the *lme* function of the *nlme* package of R (Pinheiro et al., 2023), and with paired-samples t-tests in case of the implicit transfer task, using the *rstatix* package of R (Kassambara, 2023). We conjectured that longer latencies on the miscategorized trials of the implicit transfer task and/or the self-perspective blocks of the explicit perspective-taking task would provide further evidence for the spontaneous updating of the other's mental state. While in the implicit transfer task such results would constitute alternative evidence for considering the 'partner's' different

[3] Note that the amount of data did not make it possible to use a discrete-continuous mixture (or hurdle) model, as suggested by some authors (see e.g. Ospina & Ferrari, 2010).

[4] Since the items disappeared, this was the time window when participants actually had to make their decisions how to categorize the item.

view of the previously miscategorized color, on the self-perspective trials of the explicit perspective-taking task longer latencies would essentially indicate an altercentric intrusion of the 'partner's' irrelevant perspective. In all analyses reported in the paper, tests were two-tailed, with significance level set at $p<0.05$.

2.2. Results

2.2.1. Anticipatory looking task: test phase

Analysis of the proportion of looking data in the test phase yielded no significant main effect of condition (Wald χ^2=0.22, df=1, p= .641) or block (Wald χ^2=5.27, df=3, p= .153). Importantly, however, there was a significant condition x block interaction (Wald χ^2=40.80, df=3, p< .001, for the parameter estimates see: Table S5). While the proportion of looking towards the nontarget box tended to decrease with time in the properly categorized ambiguous condition, the proportion of looking towards the target box in the miscategorized condition increased sharply (see Fig. 3a). When fitting models separately for the four blocks, we found that the effect of condition was significant from the second block on (block 1 - β=-0.06, SE=0.12, z=-0.50, p= .618; block 2 – β=0.72, SE=0.14, z=5.19, p< .001; block 3 – β=0.72, SE=0.14, z=5.21, p< .001; block 4 – β=1.07, SE=0.14, z=7.92, p< .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 color. Analysis of the first look data indicated a similar pattern: there was a sharp increase in participants' tendency to look first towards the target box on the miscategorized and a slight decrease in their tendency to look first towards the nontarget box on the properly categorized ambiguous trials, after the first block (for details see the Supplementary Materials, Section C1).

Importantly, the looking pattern was strikingly different in the two unambiguous color conditions: although there was a slight increase in the proportion of looking towards the nontarget box on those trials where the unambiguous color belonged to the same category as the miscategorized ambiguous color, resulting in a significant condition effect in the second block (β=0.30, SE=0.12, z=2.48, p= .013), this was followed by a gradual decrease in the anticipatory looking instead of the marked increase observed on the miscategorized ambiguous color trials (block 3 - β=0.22, SE=0.13, z=1.72, p= .086). Anticipatory looking remained rather low ($\leq$ 0.20) in both unambiguous conditions, throughout the task, with no significant condition effect being present in the last block (block 4 - β=0.03, SE=0.12, z=0.26, p= .792; for further details see the Supplementary Materials: Section E1).

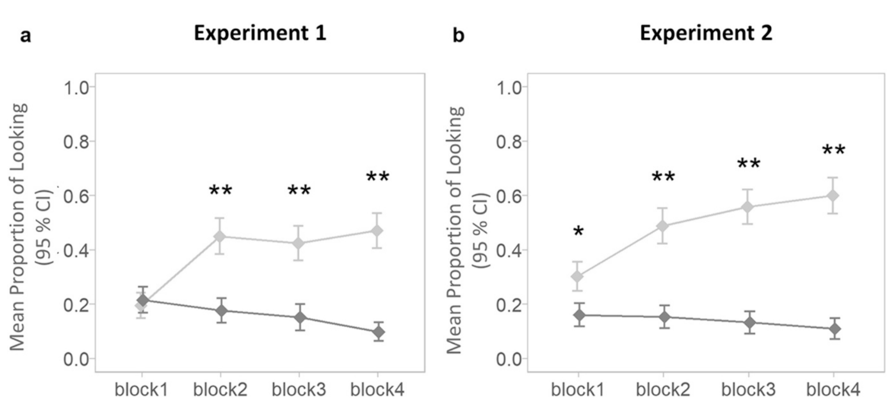
2.2.2. Implicit transfer task

The hit rate was significantly lower on the previously miscategorized than on the previously properly categorized ambiguous trials (Wilcoxon Signed Rank Test: Z=-2.32, p= .020, r=0.506, indicating a spontaneous transfer of the updated knowledge regarding how the other's sees the color that she miscategorized earlier in the anticipatory looking task. However, the difference was rather small (M_{diff}=0.12, SD_{diff}=0.28, see Fig. 4a, total sample). Inspection of the individual data showed that only 11 participants provided evidence for such an implicit transfer, by passing the figures at the gate that was an incorrect choice from self- (but correct from the other's) perspective on the previously miscategorized trials (while not making such choices on the properly categorized ambiguous trials). 2/11 participants passed the geometric figures according to the 'partner's' (updated) perspective on more than 50% of the trials, 9/11 took into account the other's perspective at least once (on 6%-50% of the trials). With respect to response times, there was no difference between the two conditions (t(29)= -0.73, p= .473, d=0.137; M_{diff}=-5.59 ms, SD_{diff}=40.65, see Fig. 4e).

2.2.3. Explicit perspective-taking task

Due to technical problems that arose during the data acquisition of 2 participants, we could analyze the data of only 28 participants (15 males, age: M_{age}= 23.64, SD_{age}=4.21).

2.2.3.1. Accuracy: geometric figures. Accuracy analysis of the figure blocks revealed a significant main effect of perspective (Wald χ^2=34.88, df=1, p< .001), but no significant main effect of condition (Wald χ^2=0.75, df=1, p= .387) or perspective x condition interaction (Wald χ^2=0.12, df=1, p= .731, for the parameter estimates see Table S14). Hit rate was generally lower when participants had to make judgments from the other's compared to their own perspective and this effect was independent of condition, suggesting that they were not entirely successful in recalling which of the two ambiguous colors their partner perceived in a different way earlier. However, this does not mean that their performance was the same in the two conditions: when categorizing the figures from the 'partner's' perspective, hit rate was lower on



	block1	block2	block3	block4
miscat.amb	0.20	0.45	0.42	0.47
propcat. amb	0.22	0.18	0.15	0.10

	block1	block2	block3	block4
miscat.amb	0.30	0.49	0.56	0.60
propcat. amb	0.16	0.15	0.13	0.11

Fig. 3. Results of the anticipatory looking task in the two experiments. Changes in the proportion of looking towards the target box in the miscategorized condition (miscat.amb: light grey line) versus the nontarget box in the properly categorized ambiguous condition (propcat.amb: dark grey line) (a) in Experiment 1 (color labels) and (b) Experiment 2 (color matching) prior the 'partner's' box selection, during the test phase of the anticipatory looking task. The figure displays the raw data. Error bars represent 95% CI. +: p<0.1; *: p<0.05, **: p<0.01.

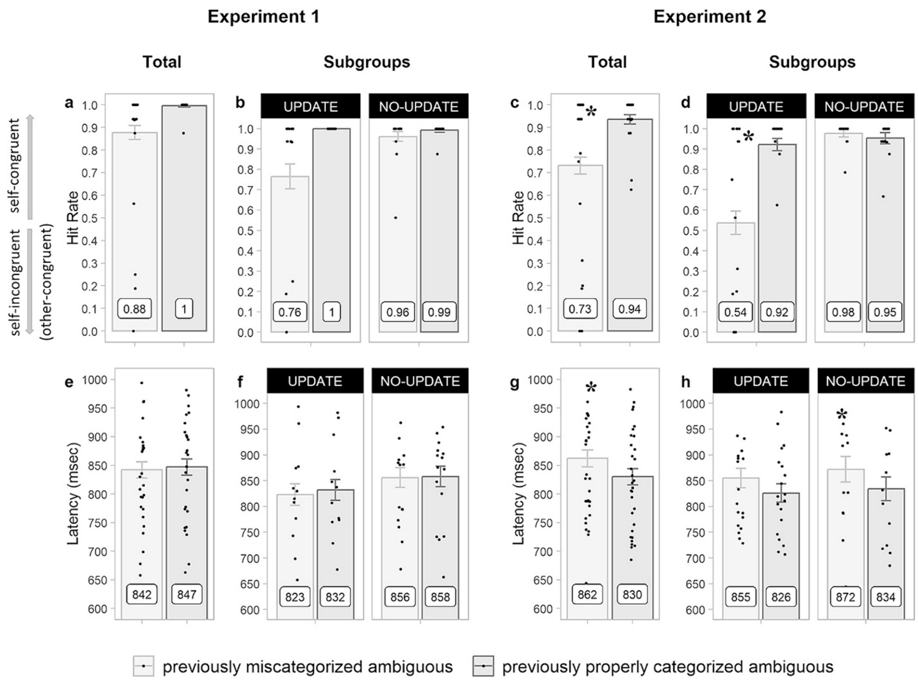


Fig. 4. Results of the implicit transfer task. The upper bar graphs depict mean hit rates while the lower ones the mean latencies in the implicit transfer tasks of the two experiments, on the previously miscategorized and previously properly categorized ambiguous trials, for the whole sample (Total), and separately for the UPDATE and NO-UPDATE subgroups. Subgroups were created on the basis of participants' performance on the explicit perspective-taking task. Passing the geometric figures incorrectly from self-perspective was coded as 0, hence, passing them according to the 'partner's' updated perspective was reflected by lower hit rates on the previously miscategorized trials. Error bars represent 95% CI. +: p<0.1; *: p<0.05, **: p<0.01.

the previously miscategorized than on the previously properly categorized ambiguous color trials (see Fig. 5a), reflecting the other-congruent (but self-incongruent) categorization in the former case, i.e. categorization that was correct from the 'partner's (updated) but not from participants' own perspective. In line with this, when fitting models separately for the two perspectives, results revealed a significant effect of condition on participants' accuracy on the other-perspective trials (miscategorized < properly categorized ambiguous: β=-0.74, *SE*=0.24, z=-3.11, *p*= .002). Importantly, there was no such significant condition effect present on the self-perspective trials (β=-0.66, *SE*=0.68, z=-0.98, *p*= .329). Such results indicate that participants were aware of the fact that the other perceived the previously miscategorized color in a different way than they themselves did and could take this into account to some extent when performing the categorization from her perspective.

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 8 took the other's (updated) perspective into account on 1 to 7 trials.

2.2.3.2. Accuracy: animal pictures. Spearman's correlation indicated a significant relationship between the hit rate of the figure and picture trials–in the previously miscategorized condition in the other-perspective blocks (r_s= .716, *p*<.001) and also on the self-perspective blocks (r_s= .600, *p*<.001). In line with this, the analysis yielded similar results for the animal pictures as for the figures: participants' accuracy was significantly lower on the other-perspective than on the self-perspective trials (perspective: Wald χ^2=36.43, df=1, *p*< .001). However, this time a significant main effect of condition also emerged (Wald χ^2=7.20, df=1, *p*= .007): hit rate was significantly lower on the previously miscategorized than on the previously properly categorized ambiguous color trials, not only when participants categorized the pictures from their 'partner's' (previously miscategorized < previously properly categorized ambiguous: β=-1.10, *SE*=0.25, z=-4.34, *p*< .001), but also when performing the categorization from their own perspective (previously miscategorized < previously properly categorized ambiguous: β=-2.01, *SE*=0.68, z=-2.94, *p*= .003). Despite the difference between the two conditions was numerically larger on the other- than on the self-perspective trials, the perspective x condition interaction was not significant (Wald χ^2=0.65, df=1, *p*= .419). Notably, such a finding likely reflects an effect of the automatically activated memories of the other's earlier actions (in relation to the miscategorized pictures) on participants' own decision process, rather than an updating of the self-perspective (as the latter would have predicted a similar effect also on the figure trials).

On the unambiguous color trials, hit rates were close to ceiling (>0.99), independent of item type, perspective and whether the color was the unambiguous version of the previously miscategorized or the previously properly categorized ambiguous color.

2.2.3.3. Latency analyses. With respect to the latencies, analyses revealed a significant main effect of perspective on both item types (figures: Wald χ^2=4.88, df=1, *p*= .027; pictures: Wald χ^2=7.14, df=1, *p*= .007), resulting from the generally longer response times on the other- compared to the self-perspective trials, but no significant main effect of condition (figures: Wald χ^2=0.02, df=1, *p*= .888; pictures: Wald χ^2=1.88, df=1, *p*= .171) or perspective x condition interaction (figures: Wald χ^2=0.26, df=1, *p*= .612; pictures: Wald χ^2=1.21, df=1, *p*= .271). It took equally long (in one case even somewhat shorter time) for participants to make their decision on the previously miscategorized and on the previously properly categorized ambiguous trials, both when taking the other's perspective (figures:β=-11.57, *SE*=15.36, *t*=-0.75, *p*= .452; pictures: β=2.41, *SE*=12.58, *t*=0.19, *p*= .848), and when performing categorization from their own perspective (figures: β=-2.33, *SE*=12.43, *t*=-0.19, *p*= .851; pictures: previously properly categorized > miscategorized - β=-16.49, *SE*=9.73, *t*=-1.70, *p*= .091), suggesting that the other's (different) perspective on the miscategorized trials did not

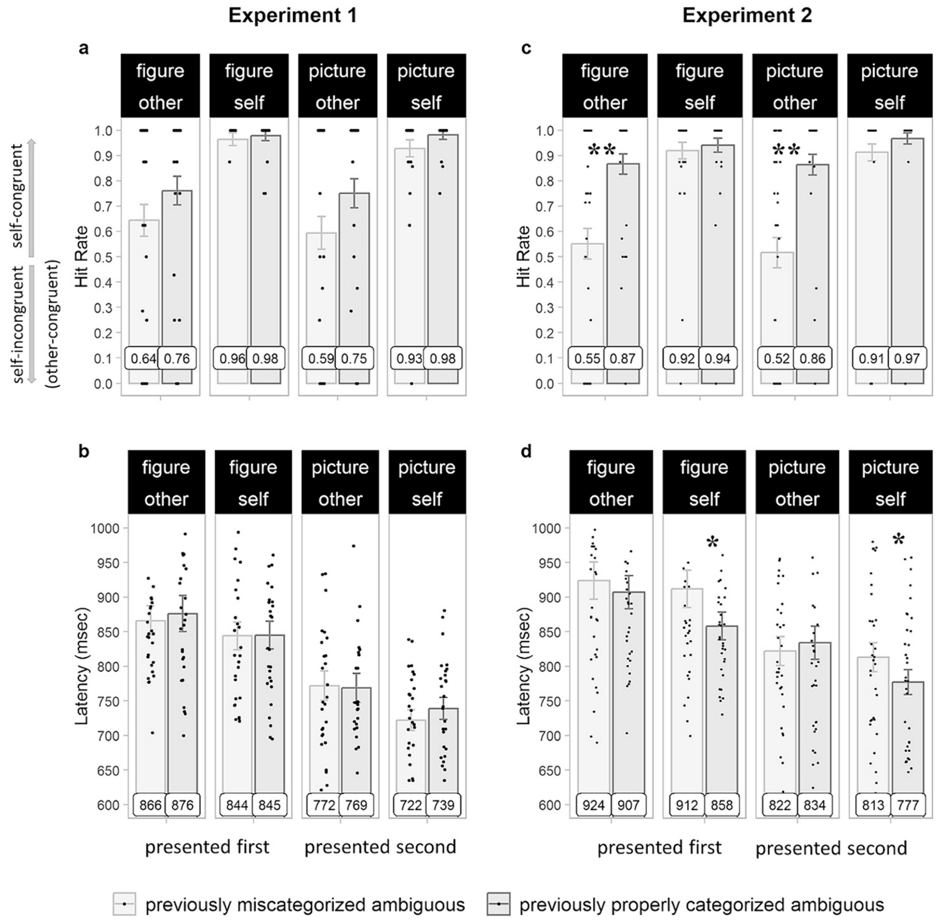


Fig. 5. Results of the explicit perspective-taking task. Mean hit rate (a, b) and latency (c, d) on the previously miscategorized and previously properly categorized ambiguous trials of the explicit perspective-taking task, in Experiment 1 and in Experiment 2. The label FIGURE denotes the geometric figures also used in the implicit transfer task, the label PICTURE the animal pictures also used in the anticipatory looking task. The labels OTHER vs SELF denote the blocks in which participants had to categorize the items from their 'partner's' versus their own perspective. The figure blocks were always presented first, for both perspectives, to ensure that categorization in the other blocks is not based on recalling how the other acted in relation to the animal pictures with the miscategorized color in the anticipatory task. Lower hit rate on the previously miscategorized trials in the OTHER blocks reflects perspective-taking. Latency refers to the time elapsed from the appearance of the cursor until the participant clicked on the figure. Error bars represent 95% CI. +: $p<0.1$; *: $p<0.05$, **: $p<0.01$.

interfere with participants' decision making.

2.2.4. Subgroup analyses based on the performance on the explicit perspective-taking task

As the large variability in participants' performance on the other trials of the explicit perspective-taking task (see Fig. 6a) suggested that participants may differ to a large extent whether or not they managed to interpret the change in their 'partner's' behavior as reflecting a change in how their 'partner' views of the miscategorized color, that is whether or not they updated the mental state of the other, next we split the sample into two subgroups, on the basis of the mean hit rate on the other-perspective trials of the explicit perspective-taking task. The 'update' subgroup consisted of participants whose responses provided evidence that they had realized that the other agent represents the miscategorized (but not the other ambiguous) color differently than they do and this was reflected their categorization of these items in the other perspective blocks: they had higher number of responses indicating perspective-taking (i.e. lower score) on the previously miscategorized than incorrect responses on the previously properly categorized ambiguous color trials (previously properly categorized ambiguous color trials hit rate – miscategorized trials hit rate > 0). The 'no-update' subgroup consisted of participants for whom there was no clear evidence for such a recomputation in the responses provided in the explicit perspective-taking task (operationalized as: previously properly categorized ambiguous – miscategorized trials hit rate $=<0$): they either did not categorize the previously miscategorized color according to the 'partner's actual perspective or were more likely to categorize the previously properly categorized ambiguous than the previously miscategorized color in a rule-incongruent manner Importantly, such a cutoff score meant that in order to be categorized as an 'updater' it was not enough for a participant to realize that their partner's perspective of the ambiguous colors differs from their own in some way (i.e., to have lower scores on the ambiguous other-perspective compared to self-perspective trials), s/he also had to be able to correctly identify which of the two ambiguous colors the other perceives in a different way than they do. This ensured that the 'update' subgroup would only include participants who had found the correct explanation for the other's unexpected behavior (a different view of a specific color) and would exclude those who might have thought that the other was simply making errors or was just uncertain in how to categorize the two ambiguous colors.

Splitting the sample in the above-mentioned way, yielded 12 participants in the 'update' (difference score > 0) and 16 in the 'no-update' subgroup (difference score≤ 0). After creating the two subgroups, we reran the main analyses separately for the two samples.

2.2.4.1. Update subgroup

2.2.4.1.1. Anticipatory looking task: test phase. Analysis run separately on the 'update' participants' proportion of looking data revealed

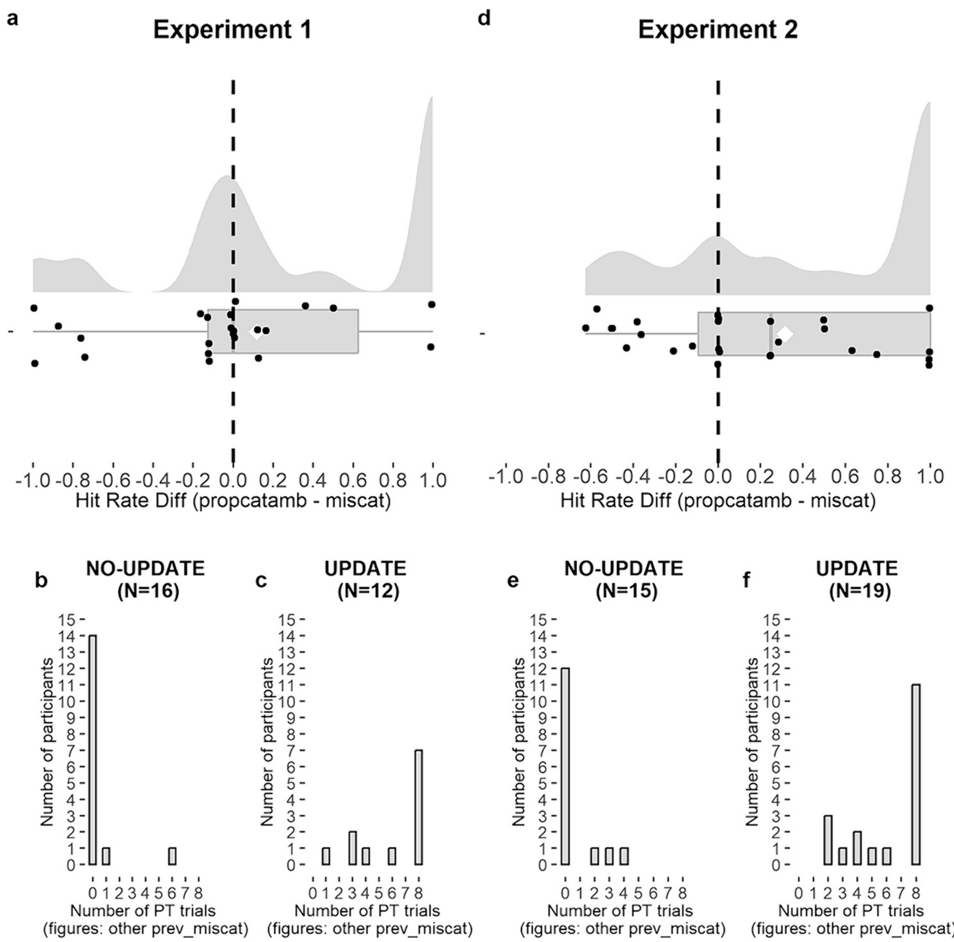


Fig. 6. Distribution of the difference scores that served as the basis of creating subgroups in (a) Experiment 1 and (d) Experiment 2. Difference scores were computed by subtracting the hit rate on the other-perspective figures previously miscategorized and previously properly categorized ambiguous color trials of the explicit perspective taking task. Those who had >0 score became members of the 'update', the rest (score ≤ 0), became members of the 'no-update' subgroup. The lower panel shows the number of trials on which participants in the respective subgroups categorized the geometric figures with the previously miscategorized color according to the 'partner's' updated perspective in the other-perspective block.

no significant main effect of condition (Wald χ^2=0.09, df=1, p= .766) or block (Wald χ^2=4.63, df=3, p= .201). There was, however, a significant condition x block interaction (Wald χ^2=41.84, df=3, p< .001), resulting from a pronounced increase in the proportion of looking towards the target box on the miscategorized trials. When fitting models separately for the four blocks, we found that, just like in the total sample, the proportion of looking (towards the target box) was significantly higher in the miscategorized than proportion of looking (towards the nontarget box) in the properly categorized ambiguous color condition from the second block on (block 1 – β=0.05, SE=0.19, z=0.28, p= .780; block 2 – β=1.29, SE=0.22, z=5.87, p< .001; block 3: β=1.34, SE=0.23, z=5.79, p< .001; block 4: β=1.98, SE=0.22, z=8.84, p< .001). Analysis of the first look data yielded very similar results (for further details see the Supplementary Materials Section C).

2.2.4.1.2. Implicit transfer and explicit perspective-taking task. In the implicit transfer task, we found a significantly lower hit rate for the previously miscategorized than for the previously properly categorized ambiguous trials (Z=-2.30, p=.022, r=0.747), indicating that participants took into account how the other views the previously miscategorized color and used what they have learnt about the other in this new task, passing the figures incorrectly from their own but correctly from their 'partner's' perspective. As can be seen from Fig. 4b, the difference was larger in this subgroup than it was 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)= -0.70, p= .497, d=0.203).

Regarding the self-perspective trials of the explicit perspective-taking task, accuracy analysis of the figures block indicated no significant effect of condition (β=-33.54, SE=99.81, z=-0.34, p= .734). There was no significant difference between the previously miscategorized and the previously properly categorized trials in terms of latencies either (β=3.64, SE=18.84, t=0.19, p= .847), suggesting that the other's different perspective did not influence these participants' own decision process. With respect to the animal pictures, mimicking our findings in the total sample, we found a significant condition effect in the accuracy (β=-2.77, SE=0.95, z=-2.92, p= .003) but not in the latency analysis (β=2.98, SE=14.46, t=0.21, p= .837). Participants made somewhat more errors when they had to categorize animal pictures with the previously miscategorized compared to the previously properly categorized ambiguous color (see Fig. 8a), possibly because the pictures activated the memory of their 'partner's' earlier actions on these items, which, in turn, interfered with participants' decision process. Notably, the short decision times could also contribute to the higher number of errors, allowing the intrusion of the irrelevant memory traces.

2.2.4.2. No-update subgroup

2.2.4.2.1. Anticipatory looking task: test phase. Analyses run on the 'no-update' participants' proportion of looking data yielded no significant main effect of the condition (Wald χ^2=1.25, df=1, p= .263), block (Wald χ^2=1.30, df=3, p= .729 or condition x block interaction (Wald χ^2=5.36, df=3, p= .147). Although 'no-update' participants tended to look somewhat more towards the target box on the miscategorized trials as the experiment unfolded (and somewhat less towards the nontarget box on the properly categorized ambiguous trials), this change in the anticipatory looking was much less pronounced than it was in the 'update' subgroup (see Fig. 6b). In line with this, follow-up analyses indicated a significant condition effect in the last block (miscategorized>properly categorized ambiguous: β=0.38, SE=0.18, z=2.10, p= .036), but not in the first three blocks (block 1 - β=-0.18, SE=0.17, z=-1.07, p= .281; block 2 - β=0.23, SE=0.19, z=-1.21, p= .225; block 3 - β=0.16, SE=0.18, z=-0.88, p= .381), implying that these participants started to revise their original expectations only by the end of the experiment, after much more evidence than 'update' participants needed. Findings were similar for the first look data (see the Supplementary Materials Section C1). To test whether the two subgroups indeed differed in how anticipatory looking evolved on the miscategorized (compared to the properly categorized ambiguous) trials, next, we reran the model for both the proportion of looking and the first look data, including the subgroup as a fixed effect (*DV* ~ *condition* * *block* * *subgroup* + *(1|id)*). Both the analyses revealed a significant condition x block x subgroup interaction (proportion of looking: Wald χ^2=12.37, df=3, p= .006; first look: Wald χ^2=8.00, df=3, p= .046, for details see the Supplementary Materials, Section B3), corroborating our findings in the separate analyses, suggesting a marked difference between the two subgroups in terms of when participants started to and to what extent did they revise their expectations.

Importantly, these results were not due to a generally lower level of anticipation in the 'no-update' subgroup: the number of trials where participants did not look at any of the boxes (proportion of 'no-anticipation' trials) did not differ between the two subgroups (Wilcoxon Rank Sum Tests: miscategorized: Z=-0.40, p= .690, r=0.080; properly categorized ambiguous: Z=-0.19, p= .851, r=0.040; for further details see Table S12).

2.2.4.2.2. Implicit transfer and explicit perspective-taking task. Regarding the implicit transfer task, results revealed that, unlike members of the 'update' subgroup, 'no-update' participants did not take into account the other's different perspective in the task: hit rate was close to ceiling on both the previously miscategorized and previously properly categorized ambiguous trials, with the analysis indicating no significant difference between the two (Z=-0.74, p=.461, r=0.249). There was no difference in terms of the latencies either (t(15)= -0.31, p= .762, d=0.077).

With respect to the self-perspective trials of the explicit perspective-taking task, just like in the 'update' subgroup, analysis of the figures block revealed no significant effect of condition on participants' accuracy (β=1.68, SE=1.11, z=1.51, p= .131). The two conditions did not differ in terms of latencies either (β=-6.96, SE=16.56, t=-0.42, p= .675): 'no-update' participants were equally slow in categorizing figures with both ambiguous colors. Regarding the animal pictures, hit rate was, again, close to ceiling on both types of trials, with no difference between the two conditions (β=-0.50, SE=1.04, z=-0.48, p= .631). The latency analysis, however, revealed a significant condition effect: surprisingly, 'no-update' participants were slower on the previously properly categorized compared to the previously miscategorized ambiguous color trials, when categorizing the these items from their own perspective (β=-31.18, SE=13.06, t=-2.39, p= .018) which may indicate an awareness that the 'partner' acted differently in relation of one of the ambiguous colors but a possible confusion regarding which of the two the other miscategorized earlier.

3. Discussion

The results of Experiment 1 indicate that participants spontaneously updated their 'partner's' perspective and revised their expectations regarding her future behavior, upon seeing her categorizing one of the colors in a different way than before. Correct anticipations started to emerge after the first few miscategorized trials of the anticipatory looking task. The amount of anticipatory looks towards the location that was congruent with the other's perspective but was incongruent with participants' own perspective, significantly exceeded what one could expect towards the task-incongruent location in a somewhat ambiguous situation, where there is some uncertainty regarding the outcome, already by the second block. The results of the explicit perspective-taking task, in specific the correct, rule-congruent categorization of the previously miscategorized color 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 beliefs regarding what categories the colors belong to (i.e., revision of the task-rule).

Importantly, 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 behavior, i.e., that they realized that their partner encoded a particular color in a different way than they themselves did. These participants 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. A closer look at the eyetracking data revealed that the effects observed at the group level were in great part driven by the looking behavior of the 'update' subgroup. although they also emerged in the 'no-update' group, by the end of the task.

Interestingly, 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 the 'update' subgroup. Despite information on how the other perceives the given, miscategorized color 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 behavior 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 of the updated perspective to a new context), the proportion of correct anticipation also remained relatively low on the miscategorized trials of the anticipatory looking task, not exceeding 0.5 in the total sample even by the last block. However, such results are not that surprising given the characteristics of our task: to update the 'partner's mental state and use this information later to predict her behavior on the one hand and to adjust to her perspective on the other, participants did not only have to overcome a general egocentric bias (i.e. how they see the color), but they also had to overwrite a presumably strong first-person color-color label association, as well as an explicit categorization rule that prescribed how the 'partner' should act upon being presented with particular colors, and which remained valid throughout the task, on the majority of the trials. These three features might have made the revision of expectations and the use of the acquired information about the other's perspective rather difficult in our experiment. To test whether this was the case, and investigate how perspective update and knowledge-transfer take place in the absence of a categorization rule, that would clearly specify which box should be selected for certain frame colors, and when no color-label associations are involved, we ran a second experiment in which the 'partner' had to judge the similarity of two colors (that of the frame and that of the boxes) instead of making categorical judgements (had to put the picture into the box s/he judged more similar in color to the picture's frame color). We assumed that such a manipulation would not only reduce the task demands but would also help participants to realize that the other's miscategorization reflects her subjective evaluation (representation) of the stimuli, which may change over time.

4. 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 color to that of the selected picture's frame. In specific, s/he was instructed to put it into the box where s/he '*thought it belongs*', on the basis of its color (into the box to which the frame's color 'resembles the most, according to his/her view'). This manipulation was expected to highlight the subjective nature of the partner's decisions, and, as a result, boost updating of the other's perspective.

4.1. Methods

4.1.1. Participants

37 students were recruited for the experiment, in the same way as in Experiment 1. The data of 3 participants were 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 anticipations in the familiarization phase of the anticipatory looking task: n=2). Thus, the final sample consisted of 34 participants (age: M= 23.82, SD=4.68, 17 males). All of them were right-handed and had normal or corrected-to-normal vision (no reported color blindness). The study was approved by the EPKEB United Ethical Committee for Research in Psychology, Hungary; participants signed informed consent and were compensated as in Experiment 1.

4.1.2. 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 color labels, blue and green colored squares were presented on the 'boxes'/'gates'. These two colors were different shades of blue/green compared to the ones used for the other stimuli but were clearly identifiable as blue/green (for an example see Fig. 1c). Second, in the anticipatory task, the 'partner' had to categorize the items according to how similar, in their view, the color of the picture frame was to that of the squares on the boxes, placing the picture into the box where they '*thought it belongs to*'.

4.2. Results

4.2.1. Anticipatory looking task: test phase

An 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 in this task, the 'partner's' actions were not based on an explicit categorization rule (see Table S1).

An analysis of the proportion of looking data in the test phase yielded a significant main effect of condition (Wald χ^2=10.49, df=1, p= .001) but significant main effect of block (Wald χ^2=1.46, df=3, p= .691) Importantly, however, was a significant condition x block interaction (Wald χ^2=32.91, 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 (for the parameter estimates see Table S5). When fitting models separately for the four blocks, we found that the effect of condition was significant already in the first block (miscategorized>properly categorized ambiguous: block 1 – β=0.37, SE=0.12, z=3.01, p= .003; block 2 – β=0.88, SE=0.13, z=6.65, p< .001; block 3 – β=1.10, SE=0.14, z=8.02, p< .001; block 4 – β=1.36, SE=0.14, z=9.59, p< .001, suggesting that participants quickly updated the mental state attributed to the other and revised their expectations regarding his/her behavior. Analysis of the first look data yielded essentially the same results: participants started to direct their first look towards the target box already in the first block on the miscategorized trials (for details see the Supplementary Materials, Section C2). Although there was a modest increase in the proportion of looking towards the nontarget box on those trials where the unambiguous color belonged to the same category as the miscategorized one, resulting in a significant condition effect by the third block (block 3 – β=0.34, SE=0.12, z=2.77, p= .006; block 4 – β=0.26, SE=0.12, z=2.15, p= .031), anticipatory looking remained rather low (≤0.22) in both unambiguous conditions, throughout the task (for further for details see: Supplementary Materials Section E2).

4.2.2. Implicit transfer task

The hit rate was significantly lower for the previously miscategorized than for the previously properly categorized ambiguous trials (Wilcoxon Signed Rank Test: Z=-2.00, p=.046, r=0.279), indicating that participants were influenced by their partner's different perspective when passing the figures (see Fig. 3c). Latency analyses revealed that it also took them significantly more time to make their decisions on these than on the previously properly categorized ambiguous trials ($t(33)$= -2.69, p= .011, d=0.462; M_{diff}=-33.69 ms, SD_{diff}=72.93, see Fig. 4d). Detailed inspection of the individual data showed that there were 15 participants who tended to choose the gate that was an incorrect choice from self- (but correct from their 'partner's') perspective on the previously miscategorized trials (reflecting correct perspective-taking), while not making such, self-incongruent choices on the properly categorized ambiguous trials. 9/15 participants passed the geometric figures according to their 'partner's' perspective on more than 50% of the trials of the previously miscategorized condition, 6/15 did so at least once during the task (on 6-50% of the trials).

4.2.3. Explicit perspective-taking task

4.2.3.1. Accuracy: geometric figures. Analysis of participants' accuracy in the figure blocks revealed a significant main effect of perspective (Wald χ^2=8.52, df=1, p= .004) but no significant main effect of condition (Wald χ^2=1.17, df=1, p= .280). There was, however, a significant perspective x condition interaction (Wald χ^2=15.49, df=1, p< .001): hit rate was significantly lower on the previously miscategorized than on the previously properly categorized ambiguous color trials when participants performed the categorization from their 'partner's' (β=-2.12, SE=0.25, z=-8.37, p< .001), but not when they categorized the figures from their own perspective (β=-0.44, SE=0.39, z=-1.14, p= .254). These results indicate that participants knew and, when instructed, could take into account that their partner perceived the previously miscategorized color in a different way than they did. The individual variation was, again, very high though: while 11/33 participants categorized the figures according to the other's (updated) perspective on all eight miscategorized trials, 12/33 did not show any sign of perspective-taking with the rest (11/33) doing so on one to seven trials.

4.2.3.2. Accuracy: animal pictures. As in Experiment 1, Spearman's correlation indicated a significant relationship between the hit rate of the figure and picture trials in the previously miscategorized condition in both the other-perspective blocks (r_s= .700, p< .001) and the self-perspective blocks (r_s= .573, p< .001). In accordance with this, the accuracy analysis of the animal pictures yielded a significant main effect of perspective (Wald χ^2=18.29, df=1, p< .001), condition (Wald χ^2=7.52, df=1, p= .006), as well as a significant perspective x condition interaction (Wald χ^2=5.96, df=1, p= .015). Participants' hit rate was lower on the previously miscategorized than on the previously properly categorized ambiguous color trials, and this difference was much more pronounced when they categorized the pictures from the other's perspective than when they performed the categorization from their own perspective (see Fig. 5c). Though the effect of condition was significant not only on the other- but also on the self-perspective trials (previously miscategorized< previously properly categorized – other: β=-2.31, SE=0.26, z=-8.99, p< .001; self: β=-1.89, SE=0.55, z=-3.43, p< .001), just like in Experiment 1, given the lack of similar effect in the figures block, this finding likely reflects an interference generated by the memory of the other's different actions and not recomputation of the first-person perspective. Hit rates were at ceiling on the unambiguous color trials ($\geq$0.99) with errors committed only on the other-perspective trials.

4.2.3.3. Latency analyses. With respect to the latencies, analyses revealed a significant main effect of perspective (figures: Wald χ^2=7.54, df=1, p= .006, pictures: Wald χ^2=18.84, df=1, p< .001) and condition for both item types (figures: Wald χ^2=7.70, df=1, p= .006; pictures: Wald χ^2=7.30, df=1, p= .007) as well as a significant perspective x condition interaction for the pictures (Wald χ^2=5.58, df=1, p= .018) but not for the figures (Wald χ^2=1.35, df=1, p= .245). When fitting models separately for the two perspectives, we found, however, similar results for the geometric figures and the animal pictures: it took significantly longer for participants to categorize the previously miscategorized items than those that were ambiguous but were previously properly categorized by their 'partner' from their own perspective (figures: β=46.95, SE=15.35, t=3.06, p= .002; pictures: β=33.85, SE=11.85, t=2.85, p= .005), implying an interference from the other's (updated) perspective . There was no such significant condition effect present for the other-perspective trials (figures: β=17.78, SE=15.94, t=1.12, p= .265; pictures β=-8.19, SE=12.52, t=-0.65, p= .513, previously miscategorized < previously properly categorized).

4.2.4. Subgroup analyses based on the performance on the explicit perspective-taking task

Splitting the sample along participants' performance on the other perspective trials, based on the same criteria as in Experiment 1, yielded 19 'update' and 15 'no-update' participants.

4.2.4.1. Update subgroup

4.2.4.1.1. Anticipatory looking task: test phase. Analysis of the proportion of looking data yielded a significant main effect of condition (Wald χ^2=20.32, df=1, p< .001) but no significant main effect of block (Wald χ^2=0.83, df=3, p= .842). There was, however, a significant condition x block interaction (Wald χ^2=26.23, df=3, p<.001). Whereas the proportion of looking towards the nontarget box did not change substantially over time on the properly categorized ambiguous trials, the proportion of looking towards the target box increased sharply from the first to the second block on the miscategorized trials and continued to increase steadily throughout the rest of the task. When fitting models separately for the four blocks, we found that, just like in the total sample, the effect of condition was significant from the first block on (miscategorized>properly categorized: block 1 – β=0.64, SE=0.16, z=4.10, p< .001; block 2 – β=1.28, SE=0.16, z=7.76, p< .001; block 3 – β=1.56, SE=0.17, z=9.03, p< .001; block 4 – β=1.71, SE=0.17, z=9.83, p< .001). Analysis of the first look data yielded highly similar results (for details see the Supplementary Materials Section C2).

4.2.4.1.2. Implicit transfer and explicit perspective-taking task. Analysis of the implicit transfer task revealed a lower hit rate for the previously miscategorized than for the previously properly categorized ambiguous trials (Z=-2.42, p= .016, r=0.567). 'Update' participants took into account the other's different perspective on roughly 50% of the trials (Fig. 3c middle panel). In fact, most of these participants (13/19) took into account the 'partner's' perspective to some degree, with only 6/19 'update' participants not doing so at all. Despite the numerically longer response latencies on the previously miscategorized trials (855 vs 826 ms), in terms of latencies, the difference between the two conditions was not significant (t(18)=-1.56, p= .136, d=0.358).

Regarding the self-perspective blocks of the explicit perspective-taking task, accuracy analysis of the figures block indicated no significant effect of condition (β=-0.70, SE=0.49, z=-1.43, p= .153). 'Update' participants were, however, significantly slower when they had to categorize figures with the previously miscategorized compared to figures with the previously properly categorized color (β=44.99, SE=19.30, t=2.22, p= .021), indicating an interference from the other's perspective. With respect to the animal pictures, we found a significant effect of condition both in the accuracy (β=-2.14, SE=0.59, z=-3.63, p< .001) and in the latency analysis (β=45.88, SE=15.25, t=3.01, p= .003): participants were not only slower but also committed more errors on the previously miscategorized than on the previously properly categorized animal pictures, which may again result from the automatic recall of the other's earlier actions and the effect of these memories on participants' own decision process.

4.2.4.2. No-update subgroup

4.2.4.2.1. Anticipatory looking task: test phase. Analysis of the test phase data revealed no significant main effect of condition (Wald χ^2=0.02, df=1, p= .890) or block (Wald χ^2=1.06, df=3, p= .787). There was, however, a tendency level condition x block interaction (Wald χ^2=6.54, df=3, p= .088). While the proportion of looking (towards the target box) started to increase in the miscategorized condition after the first few trials, the proportion of looking towards the nontarget box tended to decrease with time on the properly categorized ambiguous trials (see Fig. 7b), suggesting that, to some extent, 'no-update' participants of Experiment 2, also revised their expectations regarding their partner's behavior. Follow-up analyses revealed that the proportion of looking was significantly higher in the miscategorized than in the properly categorized ambiguous condition from the third block on (block 1 - β=0.06, SE=0.19, z=-0.33, p= .743; block 2 – β=0.28, SE=0.22, z=1.25 , p= .212; block 3: β=0.45, SE=0.23, z=1.98, p= .048; block 4: β=0.78, SE=0.24, z=3.24, p= .001). Similar findings were obtained in the first look analyses (see the Supplementary Materials Section C2).

To directly test whether the two subgroups differed statistically regarding the extent to which participants' anticipatory looking changed (i.e. how much they revised their original expectations) during the task, on the two types of ambiguous trials, just like in Experiment 1, next, we reran the proportion of looking and the first look analyses, including the subgroup as a fixed effect. The analysis of the proportion of looking data yielded a significant condition x subgroup interaction (Wald χ^2=7.34, df=1, p= .007), reflecting the marked the difference between the two subgroups in the level of anticipatory looking on the miscategorized trials, present throughout the task. Despite the difference between the two subgroups increased over the course of trials, neither the block x subgroup (Wald χ^2=0.70, df=3, p= .874), nor the condition x block x subgroup effect was significant (Wald χ^2=0.69, df=3, p= .875). There was no significant main effect of subgroup either (Wald χ^2=1.49, df=1, p= .222). The first look analyses yielded similar results: a significant condition x subgroup interaction (Wald χ^2=6.70, df=1, p= .010) but no other significant interactions with subgroup (for details see the Supplementary Materials, Section C2).

Importantly, in this experiment, the percentage of 'no-anticipation' trials was higher in the 'no-update' than in the 'update' subgroup on both the miscategorized (M_{update}=0.39, SD_{update}=0.24 vs $M_{\text{no-update}}$=0.47, $SD_{\text{no-update}}$=0.26; Z=-2.82, p= .005, r=0.49) and the properly categorized ambiguous trials (M_{update}=0.13, SD_{update}=0.17 vs $M_{\text{no-update}}$=0.45, $SD_{\text{no-update}}$=0.32; Z=-2.86, p= .004, r=0.494), i.e., the generally anticipated less than members of the 'update' subgroup, suggesting that they might have been less motivated or able to predict the other's actions, in general.

4.2.4.2.2. Implicit transfer and explicit perspective-taking task. As for the implicit transfer task, hit rates were close to ceiling on both the previously miscategorized and previously properly categorized trials, with no difference between the two conditions (Z=-1.13, p= .257, r=0.351). In fact, there were only 2/15 participants who demonstrated any sign of spontaneous perspective-taking in this task (they passed the geometric figures in line with the other's perspective on the miscategorized trials once or twice but made no errors on the properly categorized ambiguous trials). However, just like members of the

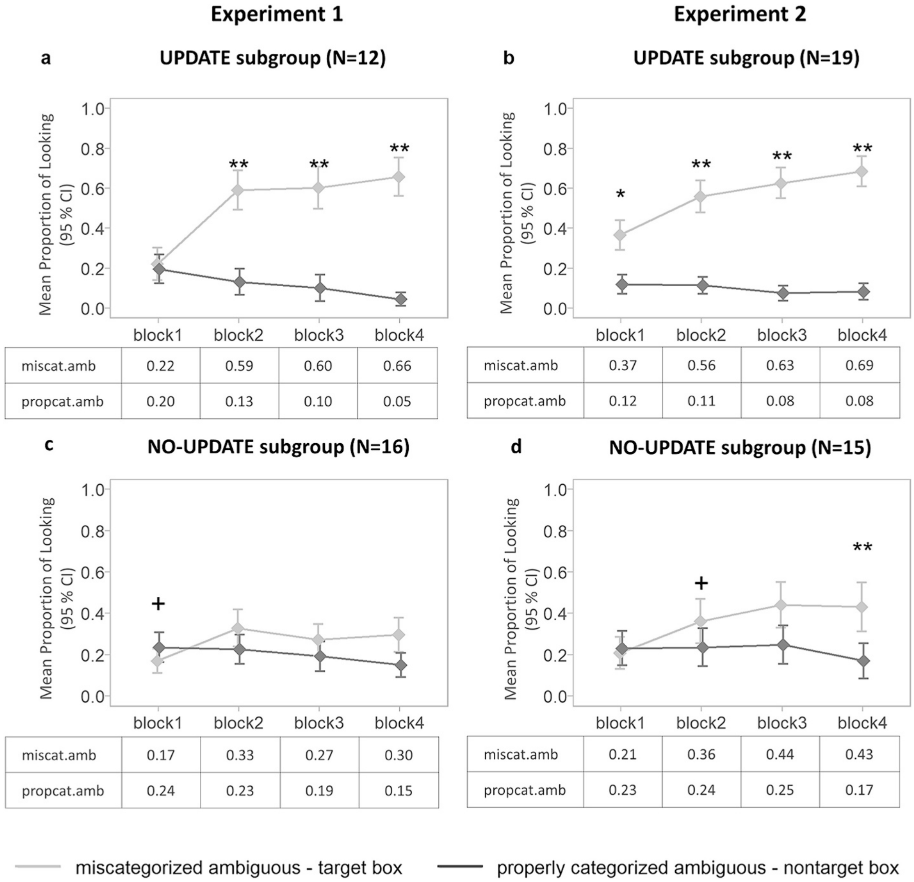


Fig. 7. Results of the anticipatory looking task in the two experiments separately for the 'update' and 'no-update' subgroups. Changes in the proportion of anticipatory looking towards the target box in the miscategorized condition (miscat.amb: light grey line) versus the nontarget box in the properly categorized ambiguous condition (propcat.amb: dark grey line) in the UPDATE (a, c) and NO-UPDATE (b, d) subgroups of Experiment 1 (color labels, left panels) and Experiment 2 (color matching, right panels), prior the 'partner's' box selection. The subgroups were created on the basis of participants' performance on the other-perspective trials of the explicit perspective-taking task. The figure displays the raw data. Error bars represent 95% CI. +: p<0.1; *: p<0.05, **: p<0.01.

'update' subgroup, 'no-update' participants were also slower on the previously miscategorized than on the previously properly categorized ambiguous trials ($t(14)= -2.44$, $p= .029$, $d=0.630$), suggesting the consideration of the other's (different) perspective.

Regarding the self-perspective block of the explicit perspective-taking task, 'no-update' participants' hit rate was at ceiling both for the previously miscategorized and for the previously properly categorized items, with analyses indicating no significant effect of condition on participants' accuracy (figures: $\beta=0.02$, $SE=0.65$, $z=0.03$, $p= .979$; pictures: $\beta=0.01$, $SE=1.42$, $z=0.01$, $p= .995$). These findings imply that the different looking pattern members of this subgroup demonstrated in the anticipatory looking task was not the result of not knowing how these colors should be categorized or the updating of the first-person perspective. Latency analysis revealed that 'no-update' participants were significantly slower when they had to categorize the figures (used in the implicit transfer task) with the previously miscategorized color compared to the figures with the previously properly categorized color ($\beta=49.84$, $SE=24.90$, $t=2.00$, $p= .047$), indicating an interference from the other's perspective. Although the pattern was similar for the animal pictures, the difference between the two conditions was smaller, resulting in no significant effect of condition for these items ($\beta=18.34$, $SE=18.68$, $t=0.98$, $p= .327$).

4.2.5. Comparison of Experiment 1 and Experiment 2

4.2.5.1. Anticipatory looking task: test phase. To investigate whether the two experiments differed significantly in terms of how anticipatory looking evolved in the two conditions over the test trials, we ran an additional model on our main dependent variable, the test phase proportion of looking data, with participants treated as random and condition, block and experiment as a fixed effects ($DV \sim condition * block * experiment + (1|id)$). Analyses revealed a tendency level main effect of experiment (Wald $\chi2=2.76$, df=1, $p= .097$) and a significant experiment x condition interaction (Wald $\chi2=7.48$, df=1, $p= .006$), reflecting the fact that in Experiment 2 anticipatory looking was generally higher on the miscategorized trials than it was 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 $\chi2=0.60$, df=3, $p= .897$) nor the experiment x condition x block interaction was significant (Wald $\chi2=1.40$, df=3, $p= .705$). When fitting models separately for the four blocks, we found a significant experiment x condition interaction in the first (Wald $\chi2=6.61$, df=1, $p= .010$) and in the last two blocks (block 2 - Wald $\chi2=1.36$, df=1, $p= .244$; block 3 - Wald $\chi2=6.68$, df=1, $p= .010$; block 4 - Wald $\chi2=6.76$, df=1, $p= .009$; for further details and parameter estimates of the models see the Supplementary Materials Section I). Comparison along the first looks yielded similar results.

4.2.5.2. Implicit transfer task. There was no significant difference between the two experiments in the extent to which participants spontaneously took into account the other's different perspective while performing the task (i.e. in terms of the difference between the previously properly categorized and miscategorized ambiguous trials' hit rate: $Z=-0.19$, $p= .852$, $r=0.025$). The two groups differed, however,

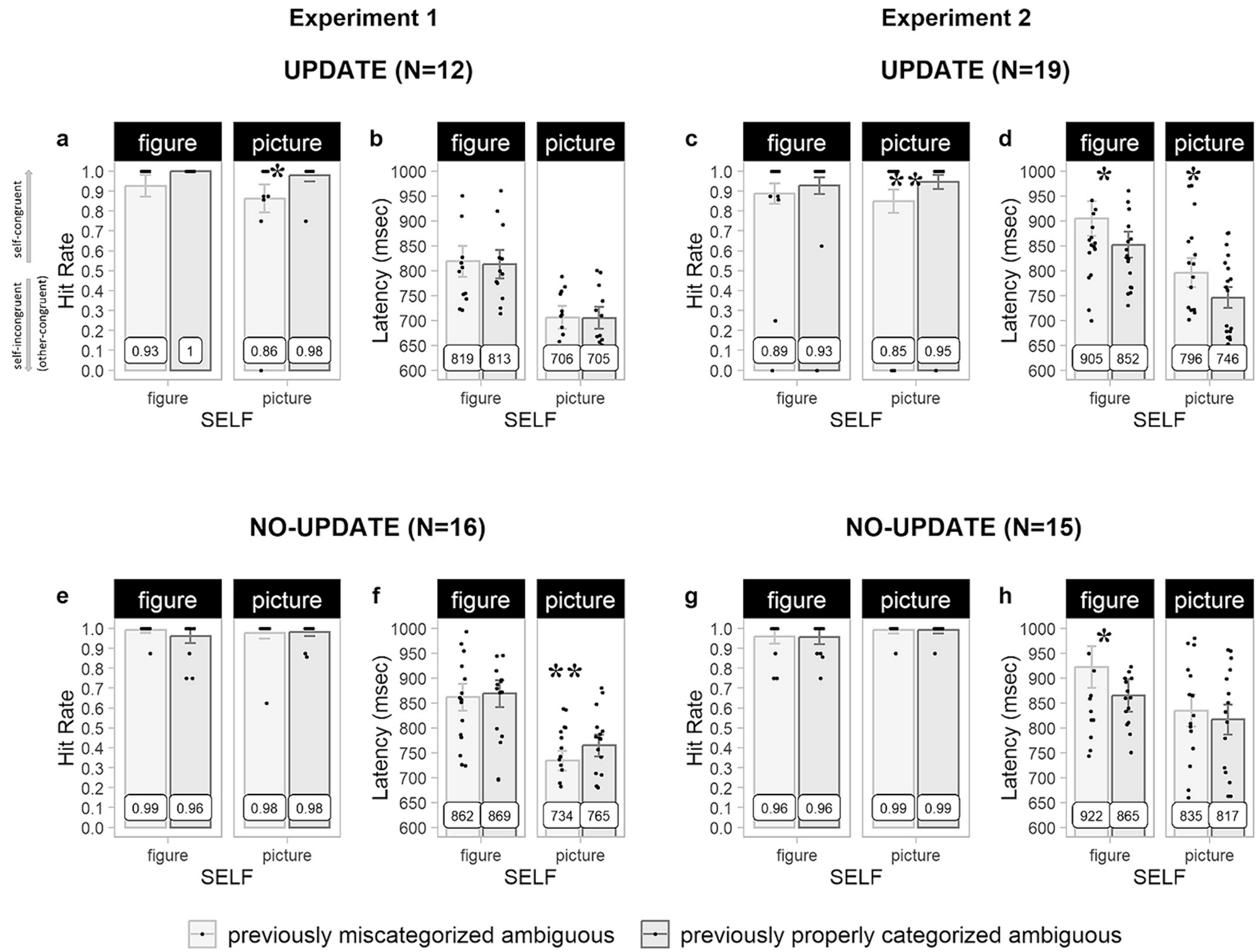


Fig. 8. Results of the explicit perspective-taking task in the self-perspective blocks of the two experiments, separately for the 'update' and 'no-update' subgroups. Mean hit rate (a, c, e, g) and mean latency (b, d, f, h) on the previously miscategorized and on the previously properly categorized ambiguous trials of in the self-perspective blocks of the explicit perspective-taking task, in the 'UPDATE' and 'NO-UPDATE' subgroups of Experiment 1 (left panel) and Experiment 2 (right panel). Subgroups were created on the basis of participants' performance on the other-perspective block of the explicit perspective-taking task. The label FIGURE denotes the geometric figures also used in the implicit transfer task, while the label PICTURE denotes the animal pictures also used in the anticipatory looking task. In the SELF perspective blocks participants had to categorize the items from their own perspective, hence it was incorrect to categorize from the other's perspective on the previously miscategorized trials Error bars represent 95% CI. +: p<0.1; *: p<0.05, **: p<0.01.

with respect to the latency of participants' responses: in Experiment 2 the difference between the previously properly categorized and miscategorized ambiguous trials' latency was significantly higher than in Experiment 1 (Welch's $t(53.3)$= -2.68, p= .010, d=0.665). Such results suggest that participants in Experiment 2 were more aware of the discrepancy between the other's perspective and their own when passing the figures to their 'partner'.

4.2.5.3. Explicit perspective-taking task. To compare the two experiments along performance on the explicit perspective-taking task we ran additional generalized linear mixed-effect models on the accuracy and linear mixed-effect models on the latency data, separately for the two item types, adding experiment as a fixed effect to the full model (*DV ~ perspective * condition * experiment +(1|id)*). Accuracy analysis of the figure blocks yielded a tendency level main effect of experiment (Wald $\chi2$=3.29, df=1, p= .069), as well as a significant experiment x perspective (Wald $\chi2$=10.90, df=1, p= .001), but no significant experiment x condition interaction (Wald $\chi2$=0.03, df=1, p= .852). There was, however, a tendency level experiment x perspective x condition interaction (Wald $\chi2$=3.79, df=1, p= .052), resulting from the higher level of perspective taking in Experiment 2 (compared to Experiment 1), i.e. lower hit rate on the previously miscategorized (than on the previously properly categorized) ambiguous color trials in the other-perspective block. In line with this, when we fitted models separately for the two perspective blocks, the condition x experiment interaction was significant for the other- (Wald $\chi2$=16.53, df=1, p< .001) but not for the self-perspective trials (Wald $\chi2$=0.05, df=1, p= .829). The pattern of results was similar for the animal pictures, with larger difference between the two experiments in the extent of perspective taking (other-perspective trials - experiment x condition: Wald $\chi2$=13.54, df=1, p< .001), resulting in a significant experiment x perspective x condition interaction for these items (Wald $\chi2$=4.88, df=1, p= .027; for further results and the exact model parameters see the Supplementary Materials Section G2).

Latency analysis of the figure blocks revealed a no significant main effect of experiment (Wald $\chi2$=0.40, df=1, p= .527). There was, however, significant experiment x condition interaction (Wald $\chi2$=4.56, df=1, p= .033), resulting from the much longer decision times of participants in Experiment 2 compared to Experiment 1, which was more pronounced on the miscategorized trials. Despite the apparent differences between the two experiments in the response time patterns on the self- versus other-perspective trials, neither the experiment x perspective (Wald $\chi2$=0.35, df=1, p= .554) nor experiment x perspective x condition interaction was significant (Wald $\chi2$=0.28, df=1, p= .600). When fitting models separately for the two perspectives, we found, however, a significant experiment x condition interaction in the self- (Wald $\chi2$=5.93, df=1, p= .015) but not in the other-perspective block (Wald $\chi2$=1.71, df=1, p= .191): participants in Experiment 2 were slower to categorize the previously miscategorized (compared to the previously properly categorized) items from their own perspective than participants in Experiment 1, suggesting a stronger interference from the other's (updated) perspective, while they did not differ with respect to the magnitude of egocentric interference present on the other-perspective trials (i.e.: of difference between the previously miscategorized versus properly categorized ambiguous condition). For the animal pictures, the experiment x perspective x condition interaction was also significant (Wald $\chi2$=5.96, df=1, p= .015), due to the markedly different response time pattern participants of Experiment 2 (previously miscategorized > previously properly categorized) demonstrated on the self-perspective trials compared to the participants of Experiment 1 (previously miscategorized < previously properly categorized; self trials - experiment x condition: Wald $\chi2$=10.13, df=1, p= .001), presumably as the result of being more successful in encoding or recalling which of the ambiguous two colors their 'partner's' miscategorized earlier, than participants were in Experiment 1 (for further details see the Supplementary Materials Section G2).

5. 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 set for the 'partner' and the presence of color labels, potentially evoking color label associations), correct anticipations tended to emerge earlier on the miscategorized trials, and were 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. In line with this, the extent of explicit perspective-taking was higher in Experiment 2 than in Experiment 1. There was, however, no significant difference between the two experiments in the number of trials on which participants spontaneously took into account how their partner views the miscategorized color when passing the figures in the implicit transfer task. This does not mean, however, that the degree to which participants considered their partner's perspective in this task was the same in the two experiments. Unlike in Experiment 1, in Experiment 2 the other's different perspective seemed to interfere with participants' responses, as indicated by the longer latencies on those trials where the presented items had the previously miscategorized color. Contrary to Experiment 1, in Experiment 2 a clear altercentric interference also emerged on the self-perspective trials of the explicit perspective taking task. Interestingly, these interference effects were present not only in the 'update' but also in the 'no-update' subgroup. This suggests that the 'no-update' participants might have been aware of the difference in the two perspectives to some extent, they just did not deploy the information they represented, either because it was not consciously accessible or because they were uncertain whether they managed to find the correct explanation for the other's behavior.

6. General discussion

The present study aimed 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 behavior that does not correspond to their original assumptions regarding the beliefs the other may hold, even when this would not be necessary in the given situation. To this end, in two experiments we investigated how participants' anticipatory looking develops when their 'partner', who performs a color categorization task, starts to categorize an intermediate, hence somewhat ambiguous color in a different way than she did before, making the updating of her mental state necessary, to be able to explain and predict her behavior. 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. In addition to the anticipatory looking task, participants were also administered an explicit perspective-taking task, in which they had to categorize colored items according to their partner's and their own perspective, to rule out non-mentalistic accounts of their anticipatory looking behavior and exclude the possibility that they updated not only the partner's but also their own perspective regarding the miscategorized color. Besides investigating how people revise their expectations about others' behavior, we also aimed to assess whether they can also use the information about the other agent's updated mental state spontaneously in subsequent social interactions. To this end, following the anticipatory looking task participants performed a virtual coordination task, in which they could apply their newly acquired knowledge about how the other sees a particular color to be more efficient in achieving a shared goal.

In both experiments, findings of the anticipatory looking task suggest that participants revised their expectations regarding the other's behavior spontaneously, without any explicit 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 explicit categorization rule and an established color-label association (Experiment 1) and also when these were not present (Experiment 2).

Importantly, in both experiments, for approximately half of the participants, performance on a subsequent explicit perspective-taking task provided clear evidence that such a change in the anticipatory looking behavior reflected the updating of the other's mental state regarding a specific color and not merely an overwriting of the previously learnt stimulus-outcome associations (pictures with blueish frames go the green box) or a more general interpretation of the other's surprising actions as reflecting an overall difficulty to correctly categorize the ambiguous colors. These participants could categorize items with the previously miscategorized color from their 'partner's' point of view, taking into account that the other's perspective differs from their own. In fact, the effects that emerged at group level in the anticipatory looking task were driven mostly, though not exclusively, by this subgroup.

One may argue that performance in the explicit perspective taking task provides only indirect evidence for the ascription of mental states the anticipatory looking task, and it does not make it possible to tell exactly what kind of mental state participants ascribed to their 'partner' (whether s/he sees the miscategorized color in a different way or has a different belief what category it belongs to). However, there are multiple reasons to believe that they encoded the *aspect* under which the other represented the previously miscategorized color, i.e. represented the different the perspective (and this way an actual mental state) of their 'partner'. First, 'update' participants could also correctly categorize items the other never acted on, from his/her perspective, suggesting that they draw a more general inference about their 'partner' in the anticipatory looking task, regarding how s/he relates to a particular color, and not just an inference about how s/he acts in relation to certain items, in a specific situation. Second, although in Experiment 1 it was possible to do this by simply learning what verbal label(s) the other associated with the colors in the anticipatory looking task, participants could not rely on such a strategy in Experiment 2. There, the only general inference they could draw was how the other encodes (and consequently acts on) the miscategorized color. Therefore, the results of Experiment 2 make it rather unlikely that our participants, in particular, the members of the 'update' subgroup, merely relied on associations (Ruffman & Perner,

2005) or on a limited ToM system that is unable represent the aspect under which an object is encoded by someone (Butterfill & Apperly, 2013), in either the anticipatory looking or the explicit perspective taking task. It is worth to note that even if participants just learned that the other agent assigns a different verbal label to the same items/features than they do in Experiment 1, this reflects an understanding of aspectuality (Perner, Mauer, & Hildenbrand, 2011), hence can be taken as evidence for genuine perspective taking. In fact, the similarity of the findings in the two experiments implies that the cognitive processes involved were, by and large, the same in Experiment 2 and in Experiment 1. Given that 'update' participants recomputed the other's mental state content, despite this was not necessary for the task they performed, these results also represent evidence that, unlike some authors suggest (Buttelmann & Buttelmann, 2017), human adults can engage in rather complex mental state computations in a spontaneous manner. In other words, they support the view that ToM can operate flexibility, yet, to a certain extent efficiently in human adults.

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 categorization rule, set for the 'partner' - "*green goes to the green box*" – that participants had to ignore 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, Bayliss, & Dux, 2012), the spontaneous updating of other agents' mental states might also be dependent on the availability of executive resources. Second, although members of the 'update' subgroup did not confuse the other's perspective with their own representation of the color in either of the two experiments, as evidenced by the low number of errors on the self-perspective trials of the explicit perspective-taking task, response latency results of Experiment 2 indicated a significant influence of the 'partner's perception of the previously miscategorized color on participants' own first-person judgments. Humans' propensity to take into account others' visual perspective spontaneously to the extent that it influences even first-person judgments (called 'altercentric intrusion effect') has been widely demonstrated in other studies, where another agent is present who holds a discrepant perspective (Samson et al., 2010; Surtees et al., 2016). Latency results from Experiment 2 suggest that such an altercentric intrusion effect 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.

Importantly, members of the 'update' subgroups also transferred what they had learnt about their 'partner' (during the anticipatory looking task) to 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 can also make use of this information spontaneously later, when interacting with the other in a new context. Crucially, this indicates that they did not just learn how their 'partner' behaves in relation to a certain color but how the other relates to it in general (i.e. what is the other's attitude towards the previously miscategorized color), as the implicit transfer task differed from the anticipatory looking task not only in terms of the presented items but also in terms of the actions participants had to perform on those.

Interestingly, however, 'update' participants acted in accordance with the other's perspective on less than 50% of the trials, passing the geometric shapes with the miscategorized color at the gates which corresponded to their own (and not to their 'partner's') perspective in most of the cases. They did so despite knowing that these items had the same color as the picture 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 relatively low even after eliminating two possible sources of interference from Experiment 1, the presence of color labels and the explicit categorization rule, set for the 'partner'. While these results might have stemmed from a strong egocentric bias and/or human adults' low propensity to generalize newly acquired knowledge about others' belief contents to subsequent social interactions, they might have also been due to certain, specific features of the task. Since participants did not see their 'partner's' actions in the implicit transfer task and were not given feedback about her success or failure, they might have thought that joint efficiency could be achieved without considering the other's different perspective when passing her the shapes (even if they select the gate that is incongruent with the other's perspective). In specific, they might have either thought that their partner is fast enough (so that it does not matter whether or not they pass the figure at the gate their partner waits) or that she would adjust to them and wait at the other gate. After all, successful coordination requires mutual adaptation from the partners and co-efficiency rarely depends on only one member of a pair.

In addition to our main results, which indicate a spontaneous updating of the other's perspective, our findings also reveal a strong egocentric bias, that was present 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 relatively low, even after eliminating the rule bias, and also in the explicit perspective-taking task, where even 'update' participants displayed some egocentric errors (roughly 20%) when judging the perspective of the other. This egocentric bias might have played an important role also in the low level of spontaneous transfer of the updated content, besides all the earlier-mentioned factors. Such results are generally in line with the findings showing that humans sometimes fail to deploy their ToM abilities in online interactions and fail to consider the other's differing visual perspective when performing a task (Keysar, Barr, Balin, & Brauner, 2000).

Interestingly, 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 color items from the other's perspective when explicitly instructed to do so and/or made a large number of errors also when categorizing the previously properly categorized ambiguous color, suggesting that they may have thought that the other is simply uncertain about which categories the two ambiguous colors belong to. It is an intriguing question how to explain the performance of this subgroup. One possibility is that members of the 'no-update' 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 or perspective, hence, might have chosen to follow a simple strategy to focus on the lighting up of the box ('their task') and not on the prediction of the other's behavior. The fact that, in Experiment 2, the number of 'no-anticipation' trials was significantly higher in this than in the 'update' subgroup supports this explanation. If this interpretation is correct and members of the no-update subgroup indeed did not engage in mentalization and, consequently, in behavioral predictions, one might argue that their results actually provide evidence for the non-automatic nature of spontaneous ToM: the computation (and consequently the recomputation) of other agents' mental state content may not be triggered under all circumstances, in everyone (see Schneider, Slaughter, & Dux, 2017).

Alternatively, 'no-update' participants might have actually started to search for a mentalistic explanation upon observing their 'partner's' unexpected actions, just did not manage to arrive at a satisfying conclusion within the available time window or arrived at it only 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 behavior 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 color. In addition, miscategorized and properly categorized ambiguous color trials were intermixed which could have made noticing the behavior change for a specific color, but not for others, in itself rather hard. 'No-update' 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 'no-update' subgroups as well, just much later than in the 'update' subgroups, lends support to this second explanation, that 'no-update' participants simply struggled more to find a proper explanation for the partner's observed behavior. This may be especially true for the 'no-update' participants of Experiment 2, who, besides starting to anticipate correctly by the end of the task, also demonstrated clear signs of interference from the other's updated perspective on the previously miscategorized trials later, in both the implicit transfer and the explicit perspective-taking task. 'No-update' participants were thus most likely 'less successful updaters', but not in fact 'no-updaters', i.e., participants who did not revise their own, original assumptions at all, regarding how the other perceives the colors. It is a matter of future research to decide whether the difficulties of this subgroup resulted from (i) the specific characteristics of our paradigm; (ii) the level of motivation no-update' participants had in the given situation to invest effort in understanding the behavior of the other, or (iii) they are related to more trait-like individual differences in the capacity to update an attributed mental state or to revise original assumptions, in general. If the observed difficulties reflect individual differences, it is a further question where those differences stem from: differences in the actual mood, the availability of executive resources or, differences in social motivation and the ability to take the perspective of another individual.

Taken together, results from Experiment 1 and 2 provide the first experimental evidence for the spontaneous updating of other agents' mental states on the basis of their observed behavior. At the same time, the findings imply that the process is by no means fully automatic and, depending on several factors, it might take shorter or longer time for a specific individual to update the content of a mental state attributed to others. The same way, individuals may also be more or less successful in transferring such knowledge to new social contexts, depending on the specifics of the situation as well as the above-mentioned trait-level individual differences. It is important note, however, that the sample size was not planned for running subgroup analyses. Therefore, the results of these analyses, should be interpreted with caution and should be considered rather as tentative evidence for existence of individual differences in the spontaneous updating of others' mental states and in the propensity/capacity of knowledge transfer. Future studies should aim to provide further support for the present findings and investigate the factors that determine the ease of updating attributed mental states and the difficulty of transferring the updated content to new contexts, using larger samples.

CRediT authorship contribution statement

Dóra Fogd: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Natalie Sebanz:** Writing – review & editing, Methodology, Conceptualization. **Ágnes Melinda Kovács:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

None.

Data availability

The data that support the findings of this study and the R scripts are openly available on the Open Science Framework and can be accessed at: https://osf.io/wka2s/.

Acknowledgments

We would like to thank Oana Stanciu for the discussions and the support provided during the revision of the paper. The study was partly supported by the European Union's Seventh Framework Programme (FP7/2007-2013) ERC Starting Grant #284236 (REPCOLLAB), and by James S McDonnell Foundation, Grant #220020449

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105812.

References

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 Behavior, 1*(4), 1–10. https://doi.org/10.1038/s41562-017-0064

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

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., … Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal, 9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

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, 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

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

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

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., … Convit, A. (2006). Introducing MASC: A Movie for the Assessment of Social Cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623–636. https://doi.org/10.1007/s10803-006-0107-0

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

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815.

Fox, J., & Monette, G. (2024). cv: Cross-Validation of Regression Models. R package version 1.1.0 https://CRAN.R-project.org/package=cv https://gmonette.github.io/cv/.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

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

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

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

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

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

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), Article 210190. https://doi.org/10.1098/rsos.210190

Kassambara, A. (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R Package Version 0.7.2. https://rpkgs.datanovia.com/rstatix/.

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

Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false beliefs in3-year-old children. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 11477–11482. https://doi.org/10.1073/pnas.1803505115

Király, I., Oláh, K., & Kovács, Á. M. (2023). Can 18-Month-Olds Revise Attributed Beliefs? *Open Mind, 7*, 435–444. https://doi.org/10.1162/opmi_a_00087

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*, 1830–1834. https://doi.org/10.1007/s13164-015-0236-5

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

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., 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), Article e0164373. https://doi.org/10.1371/journal.pone.0164373

Ospina, R., & Ferrari, S. L. P. (2010). Inflated beta distributions. *Statistical Papers, 51*, 111–126. https://doi.org/10.1007/s00362-008-0125-4

Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. *Science, 333*(6041), 474–477. https://doi.org/10.1126/science.1201216

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

Pinheiro, J., Bates, D., & R Core Team. (2023). nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-164. *https://CRAN.R-project.org/package=nlme*.

Posit Team. (2023). *RStudio: Integrated Development Environment for R*. Boston, MA: Posit Software, PBC. URL http://www.posit.co/.

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

R Core Team. (2023). _*R: A Language and Environment for Statistical Computing*_. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/>.

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

Ruffman, T., & Perner, J. (2005). Do infants really understand false belief? *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

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., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage, 101*, 268–275. https://doi.org/10.1016/j.neuroimage.2014.07.014. Epub 2014 Jul 17.

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

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

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods, 11*(1), 54.

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

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

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

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

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

van der Wel, R. P. R. D., 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