Psychological Science OnlineFirst, published on August 7, 2015 as doi:10.1177/0956797614558717

Research Article

Psychological Science
1–15

DOI: 10.1177/0956797614558717
pss.sagepub.com

A Second Look at Automatic Theory of Mind: Reconsidering Kovács, Téglás, and Endress (2010)

Jonathan Phillips[1,2], Desmond C. Ong[3], Andrew D. R. Surtees[4], Yijing Xin[5], Samantha Williams[5], Rebecca Saxe[5], and Michael C. Frank[3]

[1]Department of Psychology, Yale University; [2]Department of Philosophy, Yale University; [3]Department of Psychology, Stanford University; [4]Department of Psychology, University of Birmingham; and [5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

In recent work, Kovács, Téglás, and Endress (2010) argued that human adults automatically represented other agents' beliefs even when those beliefs were completely irrelevant to the task being performed. In a series of 13 experiments, we replicated these previous findings but demonstrated that the effects found arose from artifacts in the experimental paradigm. In particular, the critical findings demonstrating automatic belief computation were driven by inconsistencies in the timing of an attention check, and thus do not provide evidence for automatic theory of mind in adults.

Keywords

theory of mind, automaticity, false belief, replication, open data, open materials

Received 3/16/14; Revision accepted 10/16/14

Theory of mind (ToM) is the capacity to represent other agents' unobservable mental states (e.g., their goals, beliefs, or intentions) and use them in explaining or predicting their behavior and experiences. This ability is central to many aspects of human social interaction, including cooperation, moral judgments, shared attention, learning, and the ability to communicate with one another (Grice, 1989; Hare & Tomasello, 2004; Young, Cushman, Hauser, & Saxe, 2007). Within the ToM literature, the ability to represent other agents' *false* beliefs has been widely accepted as the litmus test for measuring ToM (Bennett, 1978; Dennett, 1978; Pylyshyn, 1978).

A critical theoretical question concerns the nature of the mental computations underlying ToM. One possibility is that ToM is part of central cognition, and is accordingly deliberate, slow, and effortful (Keysar, Lin, & Barr, 2003); an opposing possibility is that ToM is a modular subsystem that is automatic, fast, and effortless (Baron-Cohen, 1989; Cohen & German, 2009; Wertz & German, 2007). To understand this distinction, consider these examples: People must deliberately and effortfully multiply 17 by 18 to find that the product is 306 (multiplication is a paradigmatic central cognitive process). By contrast, they effortlessly and automatically see a point-light walker as a human body (recognition of point-light walkers is a paradigmatic modular process). Do people represent other agents' mental states as automatically as they recognize a point-light walker, or is the process more like multiplication?

The developmental literature provides some support for each of these views (for a review, see Low & Perner, 2012). If asked explicitly, children younger than 3 or 4 years old are not able to correctly predict how other people will act when their beliefs are false (Baron-Cohen, Leslie, & Frith, 1985; Wimmer, 1983). Children's eventual success on explicit false-belief tests is related to executive function and inhibitory control—key processes in regulating central cognition (Carlson, Moses, & Breton, 2002; Hala, Hug, & Henderson, 2003; Müller, Zelazo, & Imrisek, 2005). However, recent evidence suggests that even preverbal infants are able to represent other agents' false

Corresponding Author:
Jonathan Phillips, 33 Kirkland St., William James Hall, 1482, Cambridge, MA 02138-2086
E-mail: phillips01@g.harvard.edu

beliefs in simplified tasks (Knudsen & Liszkowski, 2012; Kovács, Téglás, & Endress, 2010; Luo, 2011; Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007; Surian & Geraci, 2012), broadly supporting the modular, automatic view of ToM.

In light of the developmental support for both views, a critical question is whether adults represent others' mental states automatically, or only with deliberate effort. Some evidence supports automaticity: After reading about an agent unintentionally approaching an object (e.g., approaching a drawer with perfume in it, while trying to find a hair dryer), participants wrongly endorsed mental-state-based explanations of the agent's behavior (e.g., she wanted her perfume; Wertz & German, 2007), a finding suggesting that they may have automatically computed the agent's mental state on the basis of her behavior. However, because this methodology required participants to consider the agent's mental state explicitly during the response phase, it is possible that ToM was triggered by the explicit task and not computed automatically (Back & Apperly, 2010). This work, therefore, did not provide a sufficiently rigorous test of whether adults automatically represent other agents' beliefs even when those beliefs are not relevant to, or mentioned in, the task.

Recent research by Kovács et al. (2010) was intended to provide such a test. Kovács et al. reported experiments in which the timing of adults' judgments about the presence or absence of a ball appeared to be influenced by another agent's beliefs about whether or not that ball was present, even though the agent's beliefs were irrelevant to the task. Kovács et al. used this evidence to argue that human adults automatically track other agents' false beliefs, and they connected this pattern of responses to a related demonstration of preverbal infants' false-belief representation in a similar task.

Although such a finding would be critically important to ToM research, we demonstrate that studies employing the paradigm Kovács et al. used should not be taken to inform this debate. We robustly replicated their key effects supporting automaticity of belief representation in adults (Experiment 1), but also determined that these effects arise from an artifact of the paradigm relating to the "attention check" used to ensure participants' compliance. Our conclusion is supported by three separate pieces of evidence: First, Experiments 2 through 4 show that the effects are not sensitive to the content of the agent's belief or perspective. Second, Experiments 5 and 6 show that the effects are related to the timing of the attention check in the paradigm. Third, Experiments 7 and 8 show a critical double dissociation: When the attention-check timing (but not an agent's beliefs) varies across conditions, the effect is present; when the agent's beliefs (but not the attention-check timing) vary across conditions, the effect is absent.

Taken together, these experiments provide clear evidence that the results originally offered as support for automatic ToM are better explained as the product of an unintended confound in the paradigm employed by Kovács et al. Although the experiments reported in this article do not provide conclusive evidence against the automaticity of ToM in human adults, they do strongly suggest that more research is required before any positive conclusion can be drawn.

Experiments 1–4

Method

These initial experiments stem from two independent attempts to replicate the research by Kovács et al. with two distinct and independently created sets of stimuli; these attempts were motivated by replication-based class projects (Frank & Saxe, 2012) based on the Open Science Framework's Reproducibility Project (Open Science Collaboration, 2012). The two research groups conducted Experiments 1 through 3 independently. Experiment 4 was conducted jointly by the groups.

We begin by describing the method introduced by Kovács et al. in their Experiment 1 and used with minor variations throughout our work. This task has four primary conditions (see Fig. 1 for an illustration, and see https://github.com/langcog/KTE for the stimuli). Participants are shown videos involving an agent and a ball and an occluder on a table. Following Kovács et al., we label the conditions according to whether the experimental participant (P) and the animated agent (A) believe that the ball is present behind the occluder at the end of the trial. For example, in a P+A+ trial, both participant and agent believe the ball is present behind the occluder; in a P−A− trial, neither does.

Each trial consists of viewing a video and making a judgment about the ball displayed in it. The videos start with the agent, ball, and occluder all within view on the screen. To illustrate the relatively complex paradigm used to vary the participant's and agent's beliefs independently, we explain two conditions (P+A+ and P−A+) in detail. In the P+A+ condition, the ball moves around the table; it moves behind the occluder, back into view, and finally behind the occluder again. Then, the agent leaves the scene (from the left side). While the agent is away, the ball is not visible. The agent then returns, and the occluder is lowered. At the point just before the occluder is lowered, the participant would have last seen the ball behind the occluder, and hence the participant should have a true belief that the ball is behind the occluder (P+). Similarly, the agent would have last seen the ball behind the occluder, so the agent should have a true belief that the ball is behind the occluder (A+).

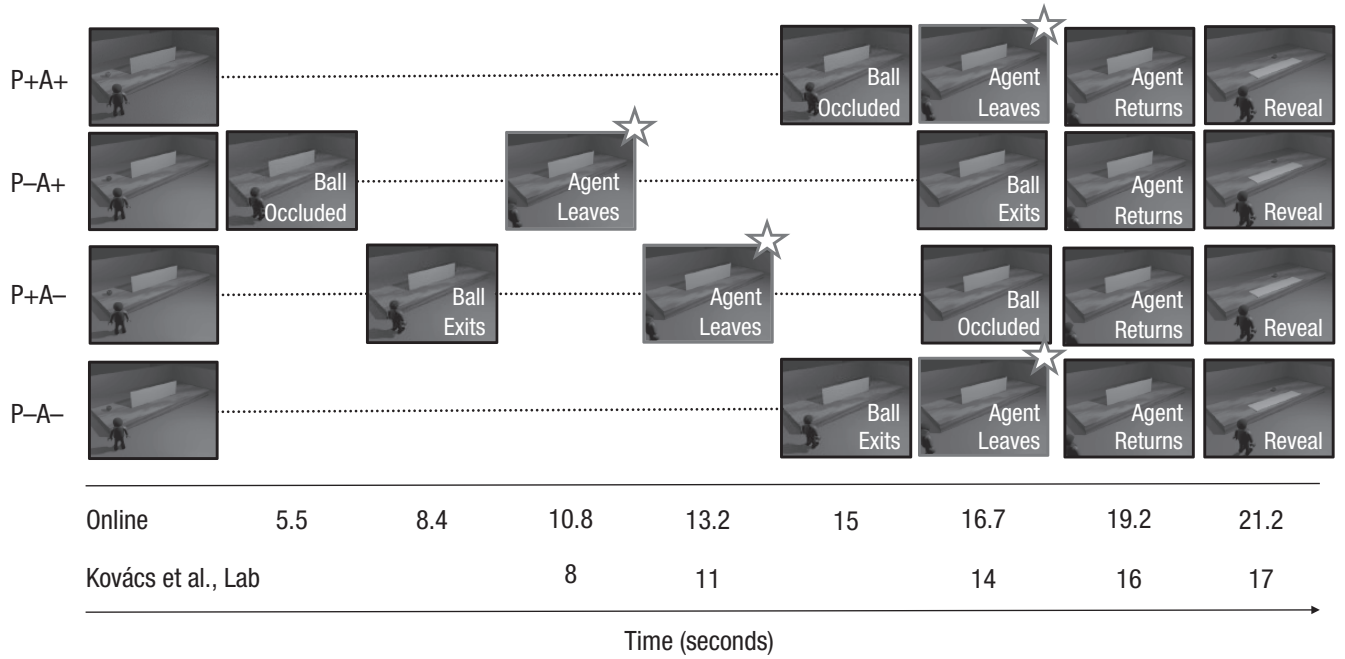

Fig. 1. Screenshots from the basic stimulus videos that were employed (with minor variations across experiments) in all experiments except Experiments 1c and 2b. In each of the four belief conditions, the video began with an animated agent standing next to a table with a ball and an occluder on it; the conditions varied in (a) whether and when the ball moved behind an occluder and then exited from behind it and (b) when the agent left the scene. In all conditions, the video ended with the agent returning and the occluder being lowered. The conditions are labeled according to whether the participant (P) and the animated agent (A) would believe that the ball is present (+) or not present (–) behind the occluder at the end of the trial. The timing of events in our videos is indicated for both the online and the in-lab experiments, as is the timing for the original videos from Kovács, Téglás, and Endress (2010). The frames in which the agent leaves the scene are highlighted by a star. In the original experiment by Kovács et al., and in our Experiments 1 through 4, these starred frames corresponded to the timing of the attention check. Additional details are provided in the text.

Now consider this same sequence of events, with one change: While the agent is offscreen, the participant sees the ball move out from behind the occluder and then offscreen (off the right side). The agent then returns to the screen from the left side. Thus, when the agent returns, the agent should have the belief that the ball is still behind the occluder (A+). But the participant has seen the ball move from behind the occluder and offscreen, and hence should believe that the ball is not behind the occluder (P–). This is the P–A+ condition. Corresponding manipulations lead to the other two complementary conditions: P+A– and P–A–. Hence, the four conditions arise from a 2 × 2 cross of whether the participant last saw the ball roll behind the occluder (P+) or move offscreen (P–), and whether the agent last saw the ball roll behind the occluder (A+) or move offscreen (A–).

On half the trials, when the occluder is lowered at the end of the video, the ball is revealed to be behind the occluder; on the other half, the ball is absent when the occluder is lowered. These two outcomes (ball present or absent) are fully crossed with the four belief conditions (i.e., eight different movies are used). With this fully crossed design, the presence of the ball when the occluder is lowered is independent of the belief of either the participant or the agent. In other words, some trials have surprising (unexpected) outcomes; this is the case, for example, when the lowering of the occluder in the P–A– condition (both the participant and the agent last saw the ball move offscreen) reveals the ball to be present.

In our experiments, participants viewed each of the eight movies five times, for a total of 40 trials. They were instructed that the experiment was a visual detection task and were asked to respond (by pressing a key) as soon as they detected the presence of a ball after the occluder fell. In all experiments except for Experiments 2a, 2b, and 3, participants were instructed to not respond if the ball was absent. The primary dependent variable was participants' reaction time (RT) in reporting the presence of the ball. Detection responses were counted only within a 3-s window.

A critical design choice made by Kovács et al. was to avoid giving a rationale for the presence of the agent, whose beliefs were completely irrelevant to the task. The agent was relevant to only one aspect of the experiment: an attention check. To make sure that participants paid

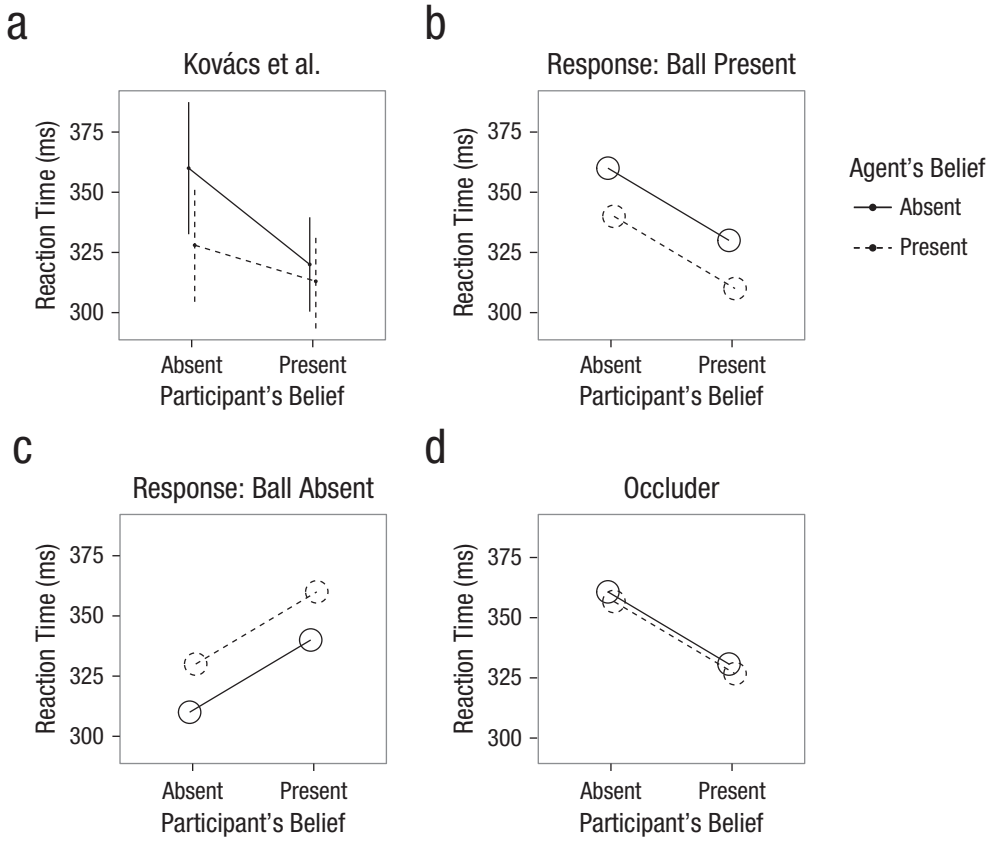

Fig. 2. Results from Experiment 1 of Kovács, Téglás, and Endress (2010) and predictions of the automatic-theory-of-mind (automatic-ToM) account. Each graph shows reaction time as a function of the participant's and agent's beliefs as to whether the ball is present or absent. The graph in (a) presents the data from Kovács et al., which have been estimated from their Figure 2A; for purposes of comparison with our other figures, the error bars show 95% confidence intervals, rather than the standard errors of the mean provided in the original. The graph in (b) depicts the pattern of reaction times predicted by the automatic-ToM account when participants are instructed to respond to the ball's presence, as in the original paradigm (tested in our Experiments 1a, 1b, and 1c). The graph in (c) depicts the pattern predicted when participants are instructed to report the ball's absence (tested in our Experiments 2a, 2b, and 3). The graph in (d) depicts the pattern predicted when participants are instructed to report the ball's presence and there is an occluder between the agent and the ball at all times (tested in our Experiment 4).

attention to the videos, participants were instructed to press a different button when the agent left the scene, which occurred at different times in the different videos. In our experiments, participants were given a 3-s window in which to respond, starting from the frame when the agent was no longer visible in the scene. Failure to respond within this 3-s window was counted as a failed attention check.

Our online experiments, conducted on Amazon Mechanical Turk, were self-paced and took an average of 20 to 25 min; the in-lab experiments took slightly less time to complete. Experiments 1c and 2b were approved by the Massachusetts Institute of Technology (MIT) Committee on the Use of Humans as Experimental Subjects; all other experiments were approved by the Stanford University Institutional Review Board. All participants gave informed consent before starting the experiment.

Predictions of an automatic-ToM account. In their original article, Kovács et al. reasoned that if participants automatically encode the agent's beliefs, and if the agent's beliefs affect RTs in reporting the presence of the ball, then participants should respond faster to the ball's presence when the agent believes the ball is present than when the agent believes the ball is absent (Fig. 2b). For comparison with this prediction, RTs in Experiment 1 of Kovács et al. are depicted in Figure 2a.

We identified two further predictions of an automatic-ToM account. First, if participants are instructed to respond to the absence, rather than the presence, of the ball, then the RT patterns should reverse (Fig. 2c).[1] For

example, when a participant is attending to the ball's absence, the fastest RTs should be observed when both the participant and the agent believe the ball to be absent and it is absent (P−A−), whereas for the original experiment, RTs are predicted to be the *slowest* in this condition. Second, if the agent's perspective is occluded (e.g., by placing a permanent occluder between the agent and the ball at all times), then the agent should not form beliefs about the ball's location and thus the agent's belief should not affect RTs for detecting the ball's presence (cf. Figs. 2b and 2d).

In Experiment 1, we attempted to replicate Kovács et al.'s evidence for the first prediction. We directly tested the second prediction in Experiments 2 and 3, and the third prediction in Experiment 4.

Stimuli. The two research groups independently created their own sets of stimuli using the description in the Supporting Online Material for the original article by Kovács et al.[2] The videos for the in-lab experiments (which were conducted at MIT) were, to the best of our knowledge, almost identical to those of Kovács et al.; the videos for the online experiments were slightly longer but contained essentially the same timing characteristics. Screenshots from the stimuli used for the four belief conditions in the online experiments, as well as the important time points in these conditions, are shown in Figure 1 (the timing of events in the videos used by Kovács et al. was estimated from watching the single sample video that was available in their Supporting Online Material at the time our experiments were conducted).

Exclusion criteria. Trials on which the attention-check response or the detection response was incorrect were dropped. In addition, participants were excluded if they failed to achieve 90% accuracy on both the attention-check and the detection response across all trials. We illustrate both of these exclusion criteria using Experiment 1a as an example. In this experiment, there were 40 trials, each with an attention-check response and a "ball present" response (or nonresponse if the ball was absent). A trial was dropped if either of the two responses for that trial was incorrect. In addition, because there were 80 responses per participant, a participant was dropped from analyses if he or she made fewer than 72 (90%) correct responses across all trials. This relatively stringent exclusion rule ensured that the data collected online through Amazon Mechanical Turk were from participants who were paying careful attention to the task, and were thus comparable to the data collected in the lab.

Note that we used the same exclusion criterion (of 90% correct responses) in all our experiments, and this decision resulted in variable exclusion rates, ranging from 30% of participants (in Experiment 4, which was online) to 0% of participants (in our in-lab experiments). We hypothesize that the exclusion rates are related to how engaged participants were in the task. For example, participants who completed the in-lab experiments were probably more motivated to stay engaged throughout the task compared with those who completed the online experiments.

Sample sizes. All target sample sizes were determined prior to data collection. For the in-lab experiments, we chose to collect data from the same number of participants as in Kovács et al. (i.e., $N = 24$). For each online experiment, we planned to collect data from a sample that was 2.5 times the size of the original (i.e., $N = 60$), which would provide 80% power to reject a detectable effect (Simonsohn, 2015). Hence, for Experiment 1a, we collected data from a sample of 60. However, because we filtered out participants with low accuracy, we subsequently decided to increase the sample size to 80 for the online experiments, to allow for adequate power even after the exclusion rule was applied. In Experiment 3, we inadvertently collected data from a sample of 100 instead.

Statistical approach. For our initial replication (Experiments 1a–1c), we followed the statistical approach used by Kovács et al., using separate *t* tests to make pairwise comparisons between conditions. However, this method both did not allow us to adequately characterize the overall pattern of results we obtained and gave rise to the concerns surrounding the use of multiple comparisons. In particular, the pattern of results that we observed was a highly consistent and characteristic crossover interaction pattern. This interaction did not conform to the pattern of data previously reported by Kovács et al., and it is clearly not predicted by their theoretical account. We discuss this crossover interaction at length when we present our results.

Because the crossover interaction described the relationship among four measurements, it could not be appropriately tested via independent *t* tests. Thus, in addition to performing *t* tests, we aggregated information across our experiments by quantifying this crossover interaction. We adopted the following summary approach: Using the lme4 package in R (Bates, Maechler, & Bolker, 2012), we fit a linear mixed-effects model (Gelman & Hill, 2007; Jaeger, 2008) with the structure "rt ~ participant.belief * agent.belief + (participant.belief * agent.belief | subject)." (Note that this model uses a "maximal" random-effects structure; Barr, Levy, Scheepers, & Tily, 2013). We then used the reliability of the crossover-interaction effect as a test of having observed the same crossover pattern as in the initial replication experiments. We report regression coefficients with their 95% confidence intervals (CIs), which were computed using the $t = z$

method because of the large amount of data collected (Barr et al., 2013). Our full models, data, and analysis code are available at https://github.com/langcog/KTE.

Experiment-specific methods. Experiment 1a (N = 60; 6 excluded) and Experiment 1b (N = 80; 8 excluded) were conducted online.[3] Experiment 1c (N = 24; 0 excluded) was conducted in person at MIT; participants were adults (17 females; ages 18–26 years, M = 20.6) tested in quiet, dark rooms. Experiments 1a through 1c were all direct replications of Experiment 1 in Kovács et al.

Experiment 2a (N = 80; 18 excluded) was conducted online. Experiment 2b (N = 24; 0 excluded) was conducted in the lab; participants were adults (21 females; ages 18–55 years, M = 22.58) tested in quiet, dark rooms at MIT. These two experiments differed from Experiment 1 in that participants were asked to press one button if the ball was present and another button if the ball was absent (i.e., a two-alternative, forced-choice response). This manipulation allowed us to measure responses to both ball-present and ball-absent trials. According to the predictions of an automatic-ToM account, participants' responses to the absence of the ball should be facilitated when the agent believes the ball is absent (just as their responses to the presence of the ball are facilitated when the agent believes the ball is present).

Experiment 3 (N = 100; 14 excluded) was conducted online; participants were asked to respond only if the ball was absent, and to not respond if the ball was present. In other words, the response criteria were exactly the opposite of those in Experiment 1, and unlike in Experiment 2, participants were not required to respond on every trial. This experiment provided an even more minimal pair in combination with Experiment 1 than did Experiments 2a and 2b.

Experiment 4 (N = 80; 23 excluded) was conducted online. The only change from Experiments 1a and 1b was that a permanent occluder was added in the videos to entirely obstruct the agent's view of the ball. This experiment tested whether RTs were sensitive to what the agent could see: If participants implicitly tracked the agent's beliefs, then the presence of this occluder, which obstructed the agent's view, would eliminate effects due to the agent's beliefs.

Results

We replicated the critical statistical results of Kovács et al. We replicated the results for the main statistical comparisons reported by Kovács et al. Table 1 reports the results of their t tests (p. 1832) and the equivalent tests for Experiments 1a through 1c. There are four main comparisons of interest. First, participants were faster to detect the ball when both the participant and the agent believed that the ball was present than when neither did (P+A+ < P−A−). Second, participants were also faster when they believed that the ball was present but the agent did not, compared with when neither they nor the agent believed that it was present (P+A− < P−A−). These first two comparisons confirm the expected result that participants' belief would have an effect on their RT. Specifically, participants were faster to detect the ball when they believed that the ball was present behind the occluder than when they expected the ball to be absent (and were presumably surprised by the presence of the ball).

Table 1. Comparison of the Results of Experiment 1 in Kovács, Téglás, and Endress (2010) and the Direct Replications of That Experiment (Experiments 1a–1c)

Comparison and experiment	t test	Cohen's d
(P–A–) – (P+A+)		
Kovács et al.	$t(23) = 3.47$, $p = .002$	0.708
Experiment 1a	$t(53) = 2.09$, $p = .042$	0.284
Experiment 1b	$t(71) = 3.33$, $p = .001$	0.393
Experiment 1c	$t(23) = 3.21$, $p = .004$	0.654
(P–A–) – (P+A–)		
Kovács et al.	$t(23) = 3.43$, $p = .002$	0.700
Experiment 1a	$t(53) = 4.49$, $p < .001$	0.611
Experiment 1b	$t(71) = 4.71$, $p < .001$	0.555
Experiment 1c	$t(23) = 3.88$, $p < .001$	0.792
(P–A–) – (P–A+)		
Kovács et al.	$t(23) = 2.42$, $p = .02$	0.494
Experiment 1a	$t(53) = 4.37$, $p < .001$	0.594
Experiment 1b	$t(71) = 4.01$, $p < .001$	0.473
Experiment 1c	$t(23) = 2.07$, $p = .05$	0.422
(P–A+) – (P+A–)		
Kovács et al.	$t(23) = 0.99$, $p = .33$	0.202
Experiment 1a	$t(53) = 0.58$, $p = .57$	0.079
Experiment 1b	$t(71) = 0.53$, $p = .60$	0.062
Experiment 1c	$t(23) = 1.59$, $p = .13$	0.324

Note: The t, df, and p values reported for Kovács et al. were taken directly from their article; the Cohen's d values were calculated from the t and df values. The conditions are labeled according to whether the experimental participant (P) and the animated agent (A) believe that the ball is present (+) or absent (–) at the end of the trial.

Third, and most important, we also replicated the critical result that Kovács et al. interpreted as providing evidence for automatic ToM: Participants were faster to respond when the agent believed that the ball was present (and the participant did not), compared with when neither believed that it was present (P–A+ < P–A–). On the basis of this comparison, Kovács et al. proposed that the agent's belief facilitated participants' detection of the ball.

Fourth, we replicated the null result that participants' RTs did not differ between the case when only the agent believed that the ball was present and the case when only the participant believed that the ball was present

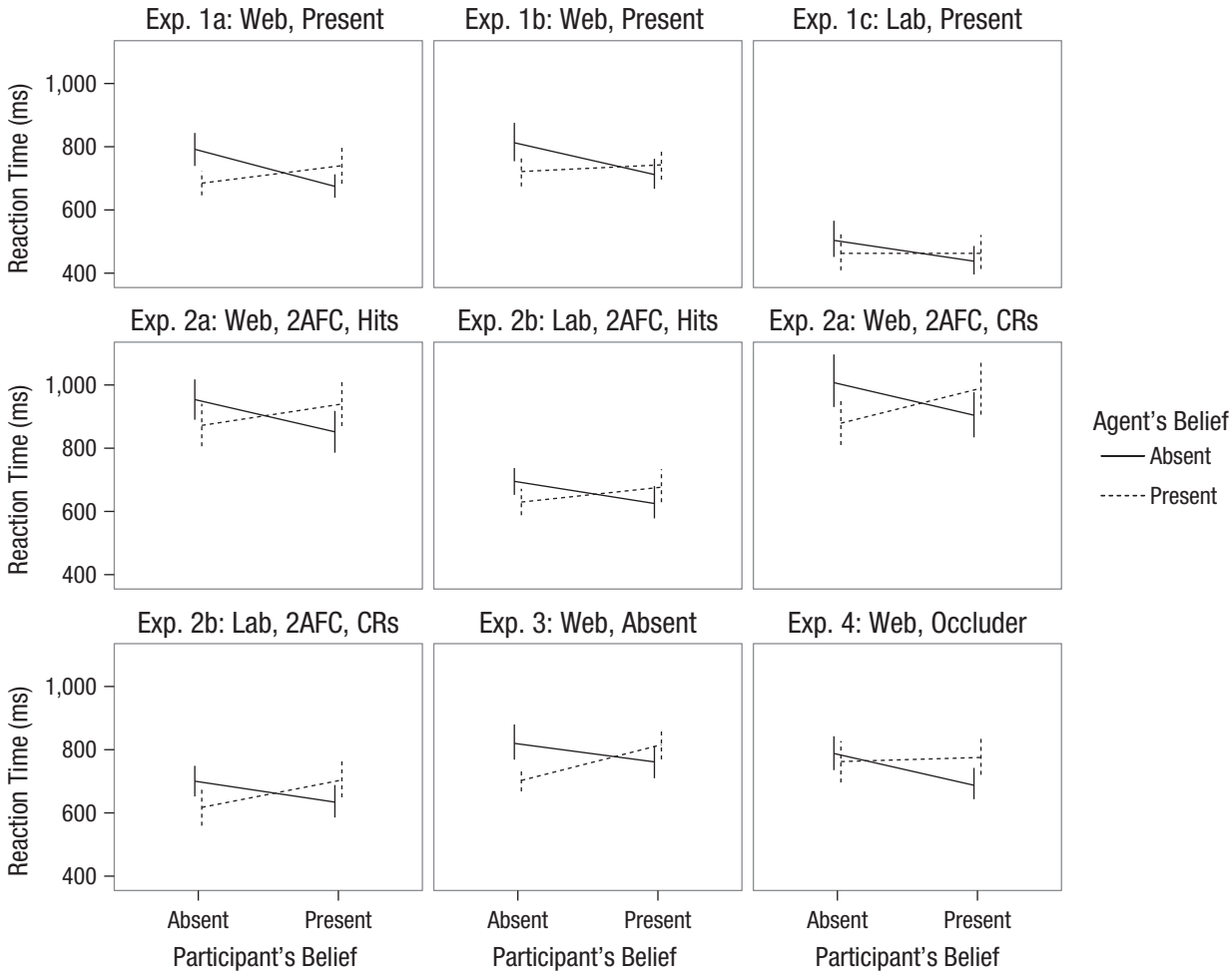

Fig. 3. Mean reaction times in Experiments 1 through 4 as a function of the participant's and agent's belief regarding whether the ball was present or absent. Error bars represent 95% confidence intervals. Lines are displaced slightly along the horizontal axis for clarity. For Experiments 1a, 1b, and 1c, which were direct replications of Experiment 1 in Kovács, Téglás, and Endress (2010), the graphs show mean reaction times for detecting the ball's presence. For Experiments 2a and 2b, in which the task was a two-alternative, forced-choice (2AFC) task, the graphs show mean reaction times for hits (responding "present" when the ball was present) and for correct rejections (CRs; responding "absent" when the ball was absent). For Experiment 3, the graph shows mean reaction times for correctly responding that the ball was absent, and for Experiment 4, the graph shows mean reaction times for detecting the ball's presence when there was a permanent occluder between the agent and the ball.

(P–A+ ~ P+A–). Kovács et al. suggested that participants' beliefs and agents' beliefs individually facilitated RTs to the same degree.

All the statistical tests that were reported by Kovács et al. were replicated in all three experiments. This robustness indicates that the effects they reported are highly replicable across different sets of stimuli and different testing environments (online vs. in the lab).

We observed a crossover interaction not consistent with automatic ToM. In addition to replicating the results that Kovács et al. reported, in Experiment 1 we observed a consistent RT pattern that they did not report: a strong crossover interaction (Fig. 3, top row).[4] The interaction coefficients were 175 ms, 95% CI = [97, 253], $p < .001$, for Experiment 1a; 121 ms, 95% CI = [65, 176], $p < .001$, for Experiment 1b; and 66 ms, 95% CI = [18, 114], $p = .007$, for Experiment 1c. The crossover was caused by relatively slow RTs on P+A+ trials. If RTs in this paradigm reflect automatic ToM, participants should be faster to respond to the ball when the agent correctly believes the ball is present than when the agent believes the ball is absent, but we observed the opposite pattern (P+A+ > P+A–; Experiment 1a: $d = 0.35$, $p = .01$; Experiment 1b: $d = 0.20$, $p = .09$; Experiment 1c: $d = 0.41$, $p = .06$. This crossover interaction is thus not consistent with automatic ToM, and it was not observed in the data that Kovács et al. reported (Fig. 1).[5] Nevertheless, this crossover interaction was robustly present in all three of these initial experiments (as well as in our subsequent experiments). Hence,

although we consistently replicated all the statistical results that Kovács et al. reported, our data are inconsistent with their theory.

The crossover interaction is observed regardless of the agent's beliefs about the presence or absence of the ball. Further evidence against interpreting the observed pattern of RTs as evidence of automatic ToM comes from Experiments 2a, 2b, and 3. Recall that the automatic-ToM account predicts that the pattern of RTs across conditions should reverse if participants are instructed to respond to the ball's absence (or, at the very least, the pattern should no longer be observed). In Experiments 2a and 2b, participants responded to both the ball's presence and its absence, and the trials of interest for this prediction are those in which participants correctly indicated that the ball was absent (i.e., correct rejections). In Experiment 3, participants responded only to the absence of the ball. The results of these experiments are shown in Figure 3.

If RTs reflect automatic ToM, participants should have been faster (or at least not slower) to respond to the absence of the ball when the agent correctly believed that the ball was absent than when the agent falsely believed that the ball was present, as illustrated in Figure 2c. Contrary to this prediction, however, participants were faster to respond to the ball's absence on P–A+ trials than on P–A– trials (P–A+ < P–A–; ds = 0.42, 0.81, and 0.66 for Experiments 2a, 2b, and 3, respectively; all ps < .001). Moreover, we observed exactly the same crossover pattern of RTs across conditions for responses to the ball's absence as we did for responses to the ball's presence (Fig. 3). The crossover interaction was significant, $p < .001$, for "absent" responses in all three experiments—Experiment 2a: $b = 207$ ms, 95% CI = [114, 301]; Experiment 2b: $b = 161$ ms, 95% CI = [102, 221]; Experiment 3: $b = 173$ ms, CI = [116, 229].

We next collapsed the data across Experiments 1 through 3 and tested whether the RT pattern for correct rejections (Experiments 2a and 2b) and "ball absent" responses (Experiment 3) differed from the pattern for hits (correct responses to the ball's presence in Experiments 2a and 2b) and "ball present" responses (Experiments 1a–1c). The model for this analysis included terms for the participant's belief, the agent's belief, whether the response was to the ball's presence or absence, and all interactions. We found a main effect of response type; RTs were overall slightly slower when the ball was absent ($b = 68$ ms, 95% CI = [39, 97], $p < .0001$). However, there were no reliable two- or three-way interactions with this term (all bs < 44 ms, all ps > .10). In addition, the two-way interaction between participant's and agent's beliefs that we observed in each individual experiment was still reliable ($b = 139$ ms, 95% CI = [105, 172], $p < .0001$). This analysis thus supports the claim that, across experiments, there was no statistical difference in the pattern of RTs across different response criteria (responding "present" or "absent"). This result clearly contradicts the predictions of an automatic-ToM account.

The crossover interaction is independent of the agent's perspective. As a final check of whether participants' RTs in this paradigm reflect automatic encoding of the agent's belief, we replicated Experiment 1 with one critical difference in the stimuli: A large wall blocked the agent's view. In this experiment, the agent had no perceptual access to the ball; thus, RTs should have been affected only by the participants' own beliefs. Yet, contra that prediction (cf. the prediction in Fig. 2d with the data in the bottom right panel of Fig. 3), the pattern of RTs across conditions remained similar to that in the previous experiments, and the crossover interaction was still reliable (interaction $b = 109$ ms, 95% CI = [49, 169], $p < .001$).

Discussion

Using two stimulus sets, we replicated the critical results of Kovács et al. in three experiments (Experiments 1a–1c) with equal or substantially greater power than the original experiment. In Experiments 2 through 4, by varying whether participants responded to the ball's presence or absence, and by eliminating the agent's perceptual access to the ball, we tested whether differences in RTs were predicted by the agent's belief about the ball. The results from these tests were inconsistent with the automatic-ToM hypothesis.

The overall pattern of RTs across conditions in Experiments 1 through 4 was better characterized by a crossover interaction than by two main effects. Note that this pattern of RTs (a) is not predicted by automatic belief tracking and (b) seems to have been driven by some between-conditions difference in the stimuli that was independent of the relationship between the agent's belief and the ball's final position. Because of the robustness of the crossover interaction, our primary goal in Experiments 5 through 8 was to try to determine what aspect of the paradigm might generate this pattern of data.

In looking for other features of the stimuli that differed between conditions, we noticed that the timing of the attention check was confounded with belief condition (see Fig. 1). Although the attention check was a minor point in the experimental design and Kovács et al. mentioned it only in their Supporting Online Material, its timing varied substantially across conditions. In order to allow for the agent's belief to differ in the different conditions, Kovács et al. manipulated the time at which the agent left the scene, which coincidentally also varied the time at which participants were required to press a

button to indicate that they were paying attention. This suggests that the timing of the attention check may have generated the pattern of RTs observed. One piece of evidence supporting this hypothesis comes from Kovács, Kühn, Gergely, Csibra, and Brass (2014), who removed the attention check and used a two-alternative, forced-choice paradigm, as in our Experiment 2. With these modifications, they observed no reliable RT differences between conditions.

We hypothesized a specific mechanism by which the attention-check confound could have generated the pattern of results we observed. When two judgments are sequential, the shorter the period of time between them, the slower the second judgment tends to be. This phenomenon is sometimes referred to as an effect of the *psychological refractory period*. Intuitively, the difficulty of making a quick response increases immediately after one has made a different quick response. Much research has investigated this phenomenon and its underlying mechanisms (e.g., Telford, 1931; reviewed in Herman & Kantowitz, 1970). In the case of the paradigm used by Kovács et al., when the attention check occurs later in the trial, participants' RTs could be slowed by the relative shortness of the "refractory period" before they are required to indicate the presence or absence of the ball. Thus, in the original design, shorter delays between the attention check and the primary response were confounded with belief condition, and the variable delays may have generated the observed pattern of results.

We tested this *attention-check hypothesis* in the next set of experiments. Experiments 5 and 6 demonstrated that the timing of the attention check plays a critical role in affecting participants' RTs. Experiments 7 and 8 went further, providing evidence of a double dissociation: The crossover effect demonstrated in Experiments 1 through 4 was present when the timing of the attention check (but not an agent's beliefs) varied across conditions, but was absent when the agent's beliefs (but not attention-check timing) varied across conditions.

Experiments 5–8

Method

The methods used in Experiments 5 through 8 were identical to those in Experiments 1 through 4 except as explicitly noted. Specifically, each trial consisted of watching a brief video of a ball and an occluder on a table and making a judgment about the ball. Participants completed 40 trials (eight movies viewed five times each) except as indicated otherwise. Participants were recruited through Amazon Mechanical Turk; the task was self-paced and took an average of 20 to 25 min. Informed consent was obtained on the first page of the experiment, and Experiments 5 through 8 were all approved by the Stanford University Institutional Review Board. These experiments were conducted collaboratively by all the authors.

Exclusion criteria. For consistency across experiments, we applied the same 90%-accuracy exclusion rule as in the first set of experiments. We again note that this decision resulted in variable exclusion rates, ranging from 30% (in Experiment 7) to 1% to 2% (in Experiments 6 and 8a). Such differences were expected because of the variation in how engaging the tasks were. For example, the videos in Experiment 7 did not show an agent but instead showed a lightbulb that flashed on at an unpredictable time; this experiment was likely both boring and difficult for participants. In contrast, the videos in Experiment 6 showed an agent, and the attention check in this experiment was predictably timed; Experiment 6 was therefore probably both easier and more engaging.

Sample sizes. Because of the exclusion criterion used to filter out participants with low accuracy, we chose a sample size of 80 for most of these experiments to allow for adequate power (i.e., 80% power to reject a detectable effect) even after exclusion (Simonsohn, 2015). In Experiments 5b and 8b, we decided a priori to increase the sample size to 200 in order to increase our power to detect higher-order interactions.

Statistical approach. We again used the coefficient on the interaction term in a linear mixed-effects model to capture the size of the crossover interaction in a way that was relatively comparable across experiments.

Experiment-specific methods. Experiment 5a ($N = 80$; 16 excluded) differed from Experiments 1a and 1b only in that the attention check was removed; that is, participants did not have to respond when the agent left the scene.

Experiment 5b ($N = 200$; 25 excluded) was designed for matched statistical comparison with Experiment 5a and Experiment 1. We planned a sample of 200 to ensure adequate power to test for the critical three-way interaction (of presence/absence of the attention-check, participant's belief, and agent's belief). In this experiment, participants completed two blocks. One block was a replication of Experiment 1 (i.e., with the attention check), and the other block was a replication of Experiment 5a (i.e., without the attention check). Block order was counterbalanced across participants. Within each block, participants viewed each of the eight videos three times; thus, there were 24 trials per block and, hence, an increased number of trials (i.e., 48).

Experiment 6 ($N = 80$; 1 excluded) differed from Experiments 1a and 1b in that the attention check was

moved to when the agent returned rather than when he left, because this event occurred at the same time (19.2 s) in all of the conditions. Thus, this experiment held the timing of the attention check constant while still ensuring participants' compliance.

In Experiment 7 ($N = 80$; 24 excluded), the agent was replaced with a stationary lightbulb that flashed on at the time when the agent would have left the scene. The lightbulb flashed on once and then stayed on for the remaining duration of the trial. Participants were instructed to respond when the light came on, as a modified attention check. This modified attention check occurred at the same times as the original attention check, but in a scenario that did not involve an agent who may have been forming beliefs about the ball's location.

In Experiment 8a ($N = 80$; 2 excluded), the agent was present (as in Experiments 1–6), and there was also a lightbulb present. As in Experiment 7, reporting when the lightbulb flashed on was the attention check. However, the time at which the lightbulb flashed on was completely dissociated from the agent's and participant's beliefs. Thus, the three attention-check timings that were present in the original videos (10.8 s, 13.2s and 16.7s) were crossed with the four belief conditions (P+A+, P+A–, P–A+, and P–A–) and the presence or absence of the ball, for a total of 24 videos. Each participant viewed each of the 24 videos twice, which resulted in an increase of the total number of trials to 48.

Experiment 8b ($N = 200$; 37 excluded) was identical to Experiment 8a except that we tested a set of five evenly spaced attention-check timings (10.9 s, 12.9 s, 14.9 s, 16.9 s, and 18.9 s), which were fully crossed with the four belief conditions and the presence or absence of the ball, for a total of 40 videos. The attention-check timings were chosen to span the range of attention-check timings we had tested earlier, from the minimum of 10.8 s to the maximum of 19.2 s. So that this experiment would not take longer to complete than the others, we had each participant view each video only once (i.e., there were 40 trials). We predicted that the lack of repetition would result in more noise and therefore chose to increase our sample size to add statistical power.

Results

The crossover interaction is observed only when there is an attention check with variable timing. In Experiment 5a, the attention-check requirement was removed, and the RT pattern became flat (Fig. 4), with no crossover interaction (interaction $b = 22$ ms, 95% CI = [–47, 91], $p = .53$).

Experiment 5b provided a replication of Experiment 1 and Experiment 5a in a within-subjects design, to allow for direct statistical comparisons between the RT patterns with and without an attention check. In this experiment, we found a reliable three-way interaction of participant's belief, agent's belief, and attention-check condition (three-way interaction $b = 76$ ms, 95% CI = [8, 145], $p = .029$). There was a crossover interaction even when there was no attention check (interaction $b = 62$ ms, 95% CI = [–16, 140], $p = .036$), but the size of the effect was more than doubled in trials with an attention check ($b = 140$ ms, 95% CI = [88, 192], $p < .001$; Fig. 4). The three-way interaction provides evidence that the magnitude of the crossover that was observed in Experiment 1 but not in Experiment 5a was driven by the attention check.

To summarize, Experiment 5 shows that removing the attention check reduces differences in RT across conditions. However, these experiments do not provide conclusive evidence for the role of the attention check; participants might simply have ignored the video display when the attention check was not required, which would have kept them from encoding either their own or the agent's beliefs.

We addressed this issue in Experiment 6 by shifting the attention check to the point when the agent returned to the scene, which was at 19.2 s in all conditions.[6] Once again, the RT pattern was flat (interaction $b = 12$ ms, 95% CI = [–35, 59], $p = .62$; Fig. 4). This experiment used the exact same stimuli as Experiments 1 through 3, except that the attention-check timing was matched across all four belief conditions, and again was based on a salient action of the agent. Critically, the characteristic pattern of RTs found in Experiments 1 through 3 was absent.[7]

In sum, Experiments 5 and 6 showed that the RT pattern observed in Experiments 1 through 4 disappeared when the attention check was removed or when its timing was held constant across all conditions, even though the stimuli were the same as those used in Experiments 1 through 3.

The pattern of observed RTs is a parametric function of the timing of the attention check and is independent of the agent's beliefs (and even the agent's presence). To directly test the attention-check hypothesis, we next decoupled the timing of the attention check from the beliefs that the participant and agent would have formed. To make this possible, we included a lightbulb in the videos and instructed participants to press a button when the lightbulb came on. This event, rather than the agent's departure, was used for the attention check. By replicating the asymmetric attention-check pattern in the absence of an agent (Experiment 7), and by varying the attention check independently of the agent's actions (Experiment 8), we were able to test for a complete dissociation between attention-check timing and belief condition.

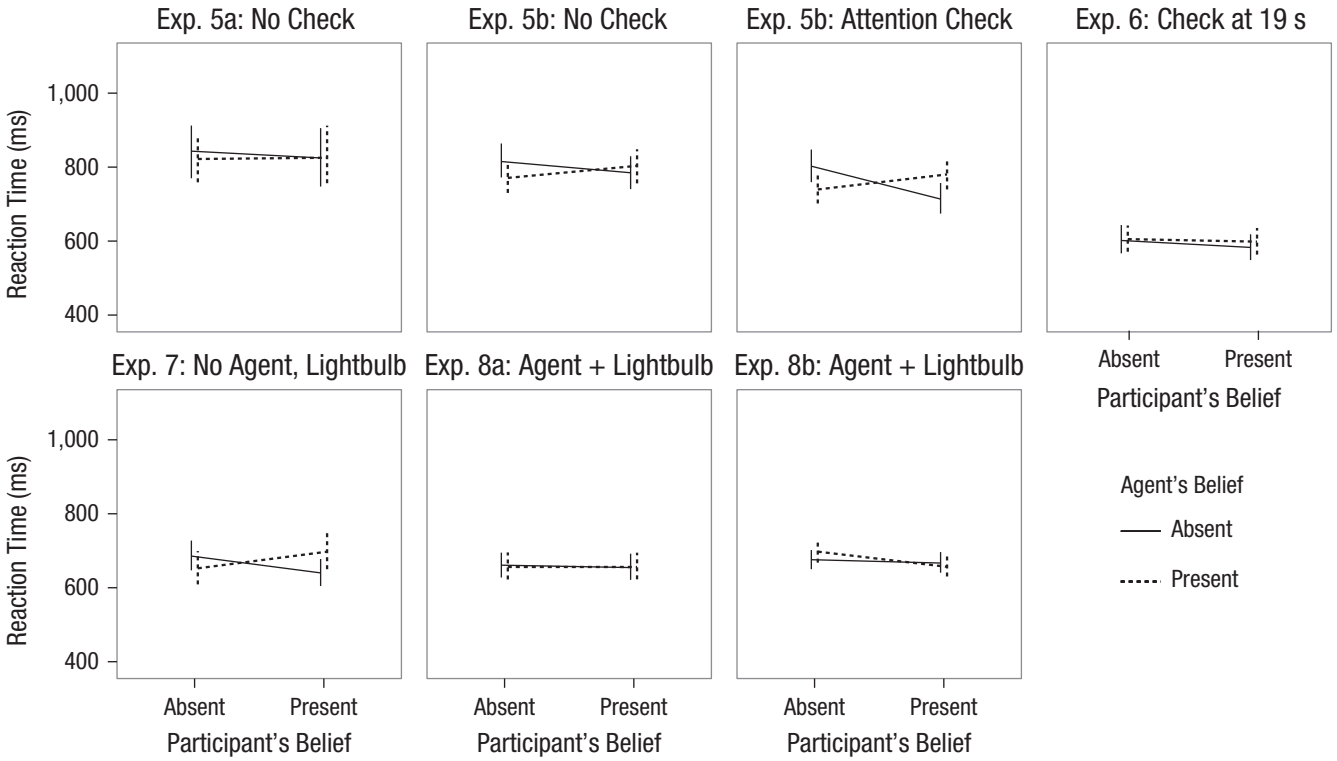

Fig. 4. Mean reaction times in Experiments 5 through 8 as a function of the participant's and agent's belief regarding whether the ball was present or absent. Error bars represent 95% confidence intervals. Lines are displaced slightly along the horizontal axis for clarity. The top row shows results for Experiment 5a, in which the attention check was removed; for Experiment 5b, in which the attention check was removed in one block and the attention check was included in another block; and for Experiment 6, in which the attention check was moved to the same time for all videos. The bottom row shows results for Experiment 7, in which the agent was removed and participants responded to the flash of a lightbulb as an attention check at the same times as in the original paradigm; for Experiment 8a, in which the agent was present but participants responded to the flash of a lightbulb at the same times as the attention check in the original paradigm; and for Experiment 8b, in which the agent was present but participants responded to the flash of a lightbulb at different, evenly spaced times.

In Experiment 7, we removed the agent entirely, but the lightbulb differentially switched on at the times that corresponded to when the agent left the scene in Experiments 1 through 4 (i.e., 10.8 s, 13.2 s, and 16.7 s; see Fig. 1.). Thus, participants were asked to respond at the exact same times in Experiment 7 as in Experiments 1 through 4. We once again observed a crossover interaction (interaction b = 86 ms, 95% CI = [32, 140], p = .002), though it was slightly smaller than before (Fig. 4). This time, however, the crossover interaction was observed without an agent being present at all! Thus, the results of Experiment 7 support the hypothesis that the RTs observed in Experiments 1 through 3 were independent of the agent's beliefs, and were plausibly driven by the attention check.

Experiment 7 showed that the RT difference between conditions can be elicited without an agent but with attention-check timings corresponding to those in the original paradigm. Experiment 8 went further by showing that even when the agent is present, the RT effect remains absent if the attention-check timing is appropriately controlled. Experiment 8a used the same timings of the lightbulb flash as Experiment 7 (10.8 s, 13.2 s, and 16.7 s), and Experiment 8b used five evenly spaced timings (10.9 s, 12.9 s, 14.9 s, 16.9 s, and 18.9 s). As in Experiment 7, participants were instructed to press a button when the lightbulb flashed on. When we averaged across attention-check timings, we found no crossover interaction in RTs in either experiment (Fig. 4; Experiment 8a: interaction b = 5.3 ms, 95% CI = [−38, 48], p = .81; Experiment 8b: interaction b = −32.6 ms, 95% CI = [−67, 2], p = .07).

To test the effect of attention-check timing on subsequent ball-detection RTs, controlling for belief condition, we added attention-check timing as a continuous predictor variable in our regression model (which fit separate coefficients for participant's and agent's beliefs and their interaction). This model showed a reliable linear effect of attention-check timing in both experiments (Experiment 8a: b = 9.7 ms/s, 95% CI = [5.5, 13.9], p < .001; Experiment 8b: b = 12.1 ms/s, 95% CI = [9.1, 15.1], p < .001; Fig. 5). The closer to the ball-detection decision the attention check was, the slower the ball-detection decision was. As

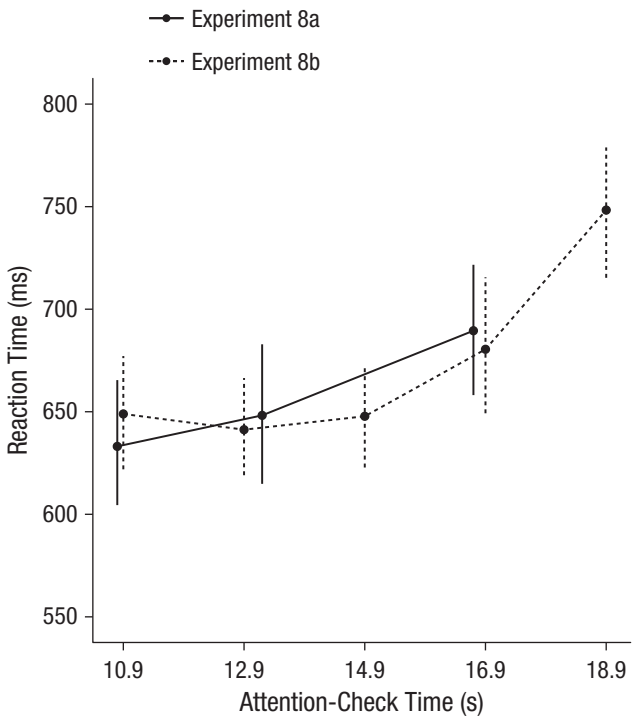

Fig. 5. Mean reaction time in Experiments 8a and 8b as a function of the timing of the attention check. Error bars represent 95% confidence intervals.

discussed earlier, this result is congruent with the literature on the psychological refractory period, which suggests that the offset between two RT measurements has systematic effects on the latency of the second measurement.

General Discussion

Collectively, these 13 experiments provide good reason to reconsider the primary evidence for automatic ToM in human adults. We began by robustly replicating the key effects that Kovács et al. claimed supported the automaticity of ToM in adults (Experiment 1). Experiments 2 through 4 then demonstrated that these effects are not sensitive to the content of the agent's beliefs or perspective. In Experiments 5 through 8, we tested the hypothesis that the key effects reported by Kovács et al. may have instead arisen from an artifact of the specific method used. Experiments 5 and 6 demonstrated that the effect was related to the presence and timing of the attention check used to ensure participants' compliance. Experiments 7 and 8 went on to show a double dissociation: The RT effects were replicated without any agent at all when the attention-check timing was asymmetric across conditions (Experiment 7), but the RT effects were completely absent even with the agent present when the attention-check timing was not confounded with the agent's beliefs (Experiment 8). In sum, the evidence across all these experiments is inconsistent with an automatic-ToM account.

These experiments also suggest one plausible mechanistic account of the attention-check timing artifact: The two conditions in which RTs were consistently highest in our replications (P+A+ and P−A−) also had the shortest delay between the attention check and the primary ball-detection response. The quick succession of the cue and the primary response likely led to increased RTs via the same refractory mechanism at play in completely nonsocial RT tasks (Herman & Kantowitz, 1970; Telford, 1931).

Stepping back to consider Experiments 1 through 8 as a whole, we conducted two random-effects meta-analyses. The first meta-analysis centered on the attention-check hypothesis and examined the evidence for the predicted crossover interaction. Figure 6a summarizes the magnitude of the crossover interaction across experiments and conditions, which are grouped by whether the attention-check hypothesis predicts a positive or null effect. The second meta-analysis focused on the automatic-ToM hypothesis and examined the evidence for the predicted effect of automatic false-belief representation (P−A+ < P−A−). Figure 6b summarizes the magnitude of this contrast across experiments and conditions, which are grouped by whether the automatic-ToM hypothesis predicts a positive or null effect. As the figure shows, the attention-check hypothesis does an excellent job of accurately explaining the results of our experiments, differentiating conditions under which we observed an effect and those under which we did not. In contrast, the automatic-ToM hypothesis does not explain the data we observed.

In conclusion, although the results Kovács et al. obtained are highly replicable, they do not provide evidence for automatic belief computation in human adults. The related evidence Kovács et al. (2010) provided for ToM in preverbal infants is, by contrast, not undermined by our experiments. Yet at the same time, our work does clearly demonstrate that the stimuli Kovács et al. used involve confounds between the agent's beliefs and the timing and sequence of critical events in the videos (see Heyes, 2014, for a related set of concerns).

It is important to keep in mind that our experiments do not provide conclusive evidence *against* automatic ToM. Rather, they highlight the need for new investigations into this aspect of human cognition. We are currently aware of only a single other study that has provided any evidence for automatic false-belief computation in human adults (van der Wel, Sebanz, & Knoblich, 2014). Given the critical theoretical importance of the question, this single study must be augmented by additional research before any positive conclusions are warranted.

Author Contributions

J. Phillips and D. C. Ong contributed equally to this work. J. Phillips, D. C. Ong, and A. D. R. Surtees, under the supervision of M. C. Frank, designed and conducted Experiments 1a, 1b, 2a, and 3 through 8; Y. Xin and S. Williams, under the supervision

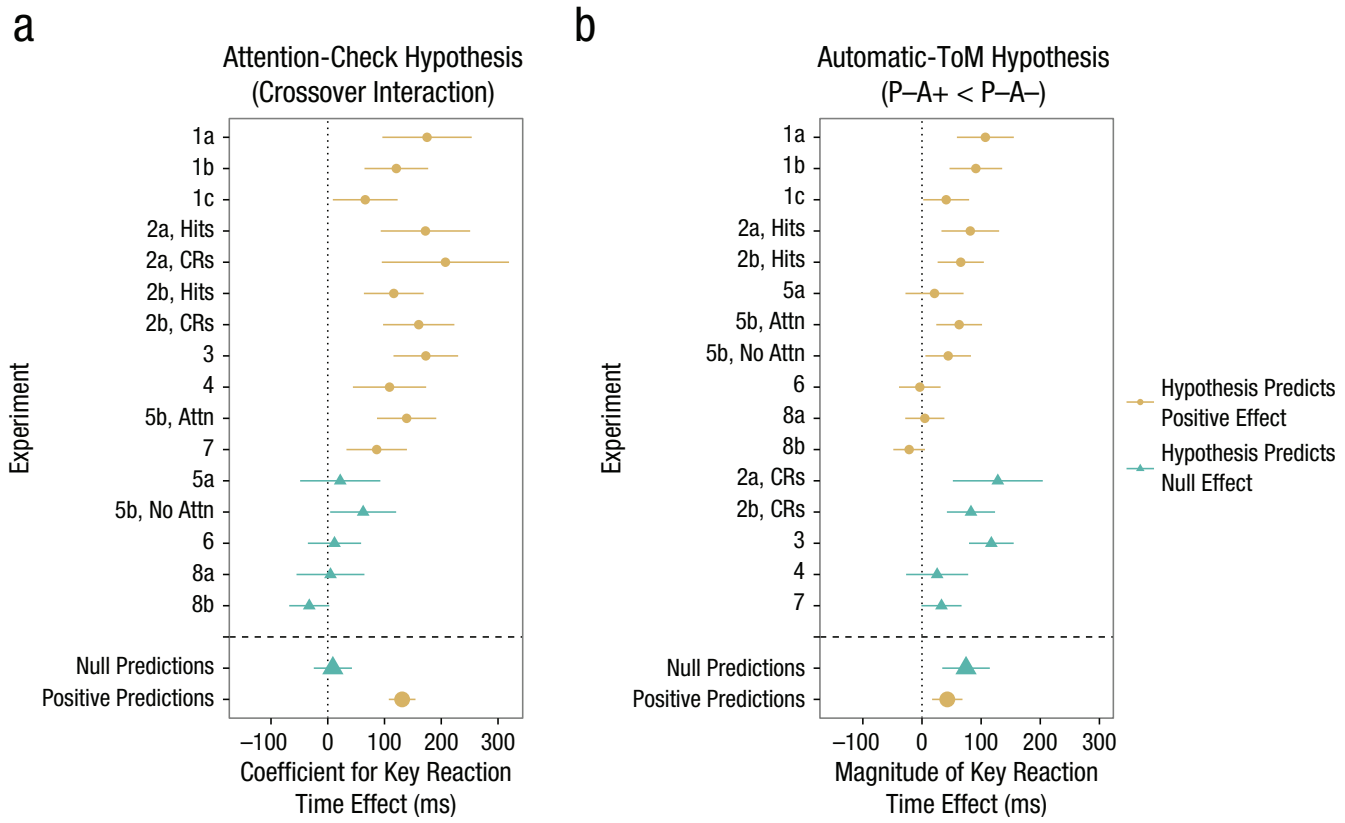

Fig. 6. Meta-analyses of the experiments reported in this article. The observed effect sizes are plotted as circles and triangles, and error bars represent 95% confidence intervals. A dotted vertical line at 0 ms is overlaid for reference. The graph in (a) presents the meta-analytic test of the attention-check hypothesis; the magnitudes of the crossover interaction are plotted. The graph in (b) presents the meta-analytic test of the automatic-theory-of-mind (automatic-ToM) hypothesis proposed by Kovács, Téglás, and Endress (2010); coefficient magnitudes are plotted for the difference between the condition in which the participant believes the ball is present and the agent believes it is absent (P–A+) and the condition in which the participant and the agent both believe the ball is absent (P–A–). In each panel, experiments and conditions are grouped according to whether or not the hypothesis predicts a positive effect; thus, the experiments are ordered differently in the two panels. The bottom two points in each panel show the meta-analytic effect sizes for the null-prediction and positive-prediction experiments, calculated using a random-effects meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010). CRs = correct rejections; Attn = attention check.

of R. Saxe, designed and conducted Experiments 1c and 2b. J. Phillips, D. C. Ong, and M. C. Frank performed the data analysis. J. Phillips and D. C. Ong drafted the manuscript, and every other author provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We would like to thank Laurie Santos and Hyo Gweon for their helpful comments on this research, as well as Jorie Koster-Hale and Alex Paunov for their assistance with the experiments conducted at MIT.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported by funding from the Stanford Psychology Department, the MIT Brian and Cognitive Sciences Department, Office of Naval Research Grant N00014-13-1-0287, and the Packard Foundation.

Open Practices

All data and materials have been made publicly available and can be accessed at https://github.com/langcog/KTE. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

Notes

1. We note, however, that beliefs about the absence of an object could be more difficult to encode than beliefs about its presence, so the pattern of RTs could be less pronounced rather than reversed entirely. Additionally, this prediction relies on the assumption that the participant's and the agent's beliefs will have an additive effect.
2. Neither group was able to obtain the original stimuli used by Kovács et al.

3. Age and gender information was not collected for the online experiments.
4. For our online experiments, RT was recorded from the first frame in which the occluder started to be lowered; the occluder took 200 ms to fall completely. From the details Kovács et al. provided in their report, it was unclear when timing began, and hence we cannot directly compare mean RTs between our experiments and theirs, as there may be constant differences between experiments in when timing began. The critical question concerns the pattern of RTs across conditions.
5. Without access to the original data of Kovács et al., we could not directly test whether our results differed reliably from theirs. Given their relatively small sample size and large CIs, however, it is possible that—although there was no crossover observed in their data—their results and ours are not inconsistent.
6. The agent returned 2 s before the occluder was lowered, so we reduced the attention-check window from 3 s to 2 s in this experiment.
7. The unusually fast overall RTs in Experiment 6 compared with the other experiments most likely arose because in this experiment, the attention check reliably appeared 2 s before the occluder fell in all conditions, so that participants were prepared to respond exactly 2 s afterward. It should not be inferred that later attention checks facilitate detection more generally; rather, the *predictability* of attention checks facilitates subsequent responding. In fact, this predictability effect is routinely found in the literature on the psychological refractory period and provides further evidence for our contention that RT measurements in the original paradigm of Kovács et al. are extremely sensitive to features of the attention check (both its relative timing and its predictability).

References

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

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry*, *30*, 285–297. doi:10.1111/j.1469-7610.1989.tb00241.x

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using 'Eigen' and S4*. Retrieved from http://CRAN.R-project.org/package=lme4

Bennett, J. (1978). Some remarks about concepts. *Behavioral & Brain Sciences*, *1*, 557–560. doi:10.1017/S0140525X00076573

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111.

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, *11*, 73–92. doi:10.1002/icd.298

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

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral & Brain Sciences*, *1*, 568. doi:10.1017/S0140525X00076664

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*, 600–604. doi:10.1177/1745691612460686

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, England: Cambridge University Press.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Hala, S., Hug, S., & Henderson, A. (2003). Executive function and false-belief understanding in preschool children: Two tasks are harder than one. *Journal of Cognition and Development*, *4*, 275–298. doi:10.1207/S15327647JCD0403_03

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, *68*, 571–581. doi:10.1016/j.anbehav.2003.11.011

Herman, L. M., & Kantowitz, B. H. (1970). The psychological refractory period effect: Only half the double-stimulation story? *Psychological Bulletin*, *73*, 74–88. doi:10.1037/h0028357

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41. doi:10.1016/S0010-0277(03)00064-7

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

Kovács, A. M., Kühn, S., Gergely, G., Csibra, G., & Brass, M. (2014). Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PLoS ONE*, *9*(9), Article e106558. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0106558

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

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology*, *30*, 1–13. doi:10.1111/j.2044-835X.2011.02074.x

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition*, *121*, 289–298. doi:10.1016/j.cognition.2011.07.011

Müller, U., Zelazo, P. D., & Imrisek, S. (2005). Executive function and children's understanding of false belief: How specific is the relation? *Cognitive Development*, *20*, 173–189. doi:10.1016/j.cogdev.2004.12.004

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

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660. doi:10.1177/1745691612462588

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *Behavioral & Brain Sciences*, *1*, 592–593. doi:10.1017/S0140525X00076895

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. doi:10.1177/0956797614567341

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

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*, *30*, 30–44. doi:10.1111/j.2044-835X.2011.02046.x

Telford, C. W. (1931). The refractory phase of voluntary and associative responses. *Journal of Experimental Psychology*, *14*, 1–36. doi:10.1037/h0073262

van der Wel, R. P., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*, 128–133. doi:10.1016/j.cognition.2013.10.004

Wertz, A. E., & German, T. C. (2007). Belief–desire reasoning in the explanation of behavior: Do actions speak louder than words? *Cognition*, *105*, 184–194. doi:10.1016/j.cognition.2006.08.002

Wimmer, H. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128. doi:10.1016/0010-0277(83)90004-5

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, USA*, *104*, 8235–8240. doi:10.1073/pnas.0701408104