Running head: RECONSIDERING AUTOMATIC THEORY OF MIND

A second look at automatic theory of mind:

Reconsidering Kovács, Téglás, and Endress (2010)

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

1 - Department of Psychology and Department of Philosophy, Yale University

2 - Department of Psychology, Stanford University

3 - Department of Psychology, University of Birmingham

4 - Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

* These authors contributed equally to this work.

Abstract

Recent research by Kovács, Téglás, & Endress (2010) argued that human adults automatically represent other agents' beliefs even when those beliefs were completely irrelevant to the task being performed. In a series of eight studies, we replicate the previous findings but demonstrate that the effects found in this work arose from artifacts in the experimental paradigm. In particular, the critical findings demonstrating automatic belief computation are driven by inconsistencies in the timing of an attention check, and thus do not provide evidence for automatic theory of mind in adults.

A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010)

"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 our ability to communicate with one another (Hare & Tomasello, 2004; Young, Cushman, Hauser & Saxe, 2007; Grice, 1989). Within ToM research, 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 illustrate, people must deliberately and effortfully multiply 17 x 18 to find that the product is 306 (a paradigmatic central cognitive process). By contrast, they effortlessly and automatically see a point-light walker as a human body (a paradigmatic modular process). Do we represent other agents' mental states as automatically as we 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 three or four years are not able to correctly predict how others will act when their beliefs are false (Baron-Cohen et al., 1985; Wimmer, 1983). Children's eventual success on the explicit false belief test 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). By contrast, recent evidence suggests that even preverbal infants are able to represent other agents' false beliefs in simplified tasks (Onishi & Baillargeon, 2005; Knudsen & Liszkowski, 2012; Kovács, Téglás, & Endress, 2010; Luo, 2011; 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., because she wanted her perfume; Wertz & German, 2007), suggesting that participants may have automatically computed the agent's mental states based on the agent's behavior. However, because this methodology requires participants to consider the agent's mental states explicitly during the response phase, it is possible that ToM is triggered by the explicit task and not computed automatically (Back & Apperly, 2010). More evidence is needed, therefore, to 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, Téglás, and Endress (2010; KTE hereafter) aimed to provide such a test. KTE 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. KTE used this evidence to argue that human adults automatically track other agents' false beliefs, and

connected this pattern of responses to a related demonstration of preverbal infants' false belief representation in a similar task.

While such a finding would be critically important to ToM research, we demonstrate that studies employing KTE's paradigm should not be taken to inform this debate. We robustly replicate KTE's key effects supporting automaticity of belief representation in adults (Studies 1a–1c), but also show that this effect arises from an artifact of the paradigm relating to the "attention check" used to ensure participant compliance.

Our conclusion is supported by three separate pieces of evidence: (1) Studies 2–4 show that the effect is not sensitive to the content of the agent's belief or perspective, (2) Studies 5–6 show that the effect is related to the timing of an attention check in the paradigm, and (3) Studies 7–8 show a critical double dissociation: when attention check timings (but not the agents' beliefs) vary, the effect is present; when the agent's beliefs (but not the attention check timings) vary, the effect is absent.

Taken together, these studies provide clear evidence that the results originally offered as support for automatic theory of mind are better explained as the product of an unintended confound in the paradigm employed by KTE. While the studies reported in this paper 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.

Study 1–4 Methods

General description. The four initial studies reported in this paper stem from two independent attempts to replicate KTE (2010) 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 project (Open Science Framework, 2012). The basic set of replications (Studies 1–3) were done independently by the two groups. Study 4 was jointly conducted by both groups.

We begin by describing the method introduced by KTE 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 sample stimuli and experiments). In each experiment, participants were shown videos that consist of an agent, and a ball and an occluder on a table. Following KTE, conditions in the experiment are notated by 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 a ball displayed in the video. The videos start with the agent, ball, and occluder all within view on the screen. To illustrate the relatively complex paradigm used to vary both the participant's and agent's beliefs independently, we explain two conditions (P+A+ and P-A+) in detail. In the first (P+A+) condition, the ball moves around the table, moving behind the occluder, moving into view, and finally moving 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 had last seen the ball behind the occluder, so the agent should have a true belief that the ball is behind the occluder ("A+").

Now consider this same sequence of events, with one change: while the agent is off-screen, the participant sees the ball moves out from behind the occluder, and then off-screen (off the right side). The agent then returns to the screen from the left side. Thus, when the agent returns, the agent still should have the belief that the ball is still behind the occluder ("A+"). But the participant saw the ball move from behind the occluder and off-screen, 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 combinations involve a 2x2 cross of whether the participant last saw the ball roll behind the occluder (P+) or move off-screen (P-), and whether the agent last saw the ball roll behind the occluder (A+) or move off-screen (A-).

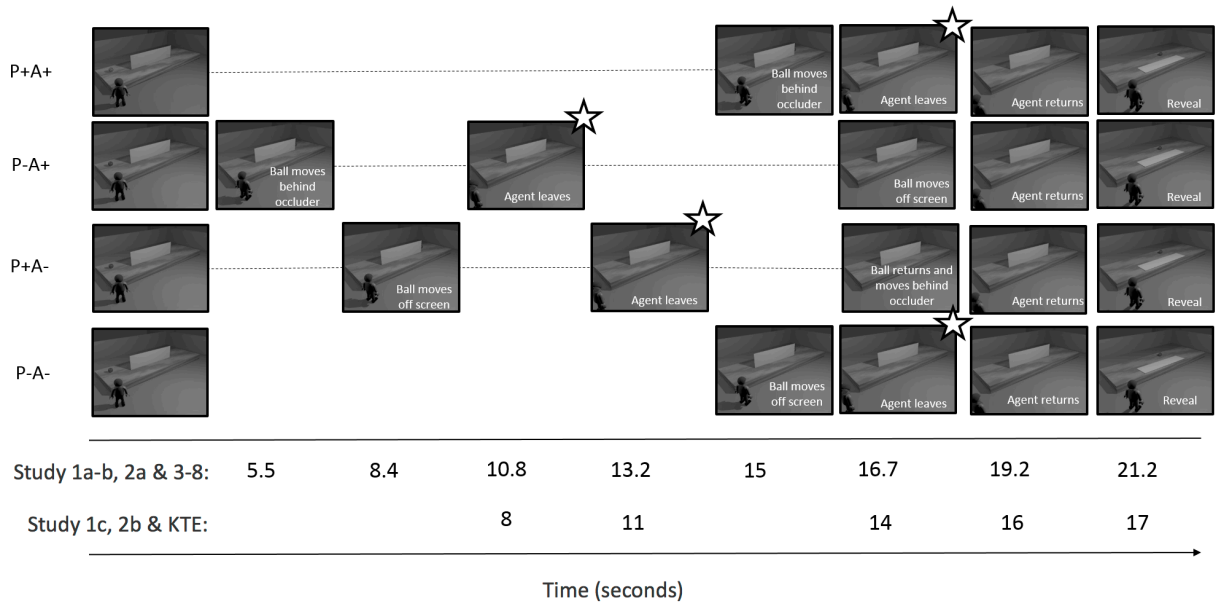

Figure 1. Screenshots (from the stimuli used in all studies except 1c and 2b) showing the sequence of events in each of the four belief conditions. The timings for our stimuli (and KTE's stimuli) are indicated. The frames where the agent leaves the scene are indicated by a star. In the original KTE study, and in Studies 1–4, these starred frames corresponded to the attention check demanded of the participants.

On all trials, after the events described above, the occluder lowered. On half of the trials, the ball was revealed to be behind the occluder; on the other half, it was absent. These two outcomes (ball present or absent) were fully crossed with the 4 belief combinations above, resulting in 8 different movies. With this fully crossed design, the presence of the ball when the occluder was lowered was *independent* of the beliefs of either the participant or the agent. In other words, there are trials with surprising (unexpected) outcomes where, for example, both the participant and the agent last saw the ball move off-screen (P-A-), but when the occluder was lowered, the ball was present.

Participants viewed each of the 8 movies 5 times, for a total of 40 trials. They were instructed that the experiment was a visual detection task, and were asked to respond as soon as they detected the presence of a ball but not to respond if the ball was absent. The primary dependent variable in KTE is participants' response time in detecting the presence of the ball. Detection responses were only counted within a 3s window.

A critical design choice in KTE's paradigm was to avoid giving a rationale for the presence of the agent, whose beliefs were completely irrelevant to the task. The agent was only relevant to one aspect of the experiment: as an attention check. To make sure that participants paid attention to the video, 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 studies, participants were given a 3s window in which to respond, starting from the frame when the agent was no longer visible in the scene. Failure to respond within this 3s window was counted as a failed attention check.

Our online studies, conducted on Amazon Mechanical Turk, were self-paced and took an average of 20–25 minutes. Informed consent was obtained on the first page of the experiment.

Studies 1c and 2b were approved by the MIT Committee on the Use of Humans as Experimental Subjects; all other studies were approved by the Stanford University Institutional Review Board.

Predictions of an automatic ToM account. Reaction times in KTE's Experiment 1, estimated from their Fig. 2A, are depicted in the leftmost panel of Fig. 2. In their original paper, KTE predicted that: if participants automatically encode the agent's belief, and if the agent's beliefs affect response times to the presence of the ball, then participants should respond faster to the ball whenever the agent believes the ball is present than whenever the agent believes the ball is absent (Fig. 2 top: 2nd panel from left).

We identified two further predictions of an "Automatic ToM" account of these results. First, if participants are instructed to respond to the *absence* of the ball, rather than the presence of the ball, then we should see a reversal of the response time patterns (Fig. 2 top: 3rd panel from left).[1] To illustrate this prediction: When attending to ball absence, the fastest reaction times should come in trials where both the participant and the agent believe the ball to be absent and it is absent (P-A-), whereas for the original experiment this was predicted to be the *slowest* 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 should not affect the response times on ball present trials (Fig. 2; compare 2nd panel from left with right-most panel).

[1] We note, however, that beliefs about the "absence" of an object could be comparatively more difficult to encode, so the pattern of reaction times could be mitigated rather than reversed entirely. Additionally, this account of KTE's predictions relies on the assumption that the participant's and the agent's beliefs will have an additive effect.

Studies 1a–1c attempt to replicate KTE's evidence for the first prediction. In addition, we directly test the second prediction in Studies 2–3, and the third prediction in Study 4.

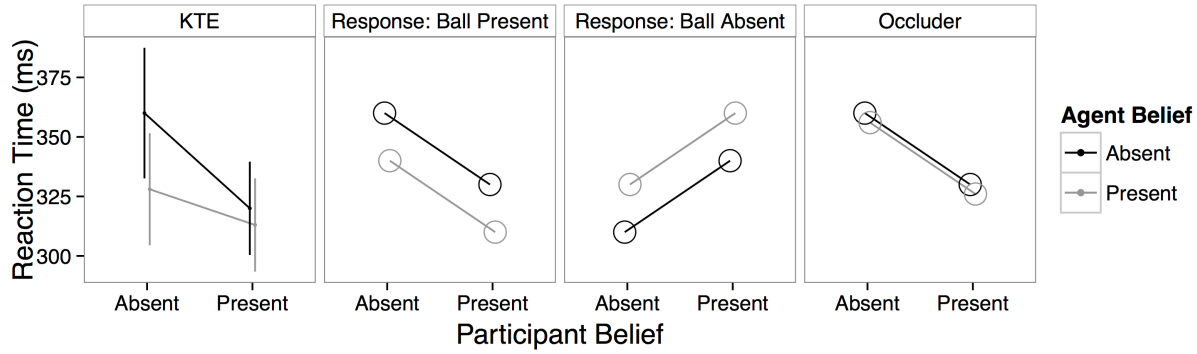

Figure 2: Left-most panel: data from KTE's Experiment 1, estimated from KTE's Fig. 2A. For purpose of comparison to other figures, error bars show 95% confidence intervals, rather than the standard error of the mean provided in the original. The next three panels show the pattern of reaction times that are predicted by Automatic ToM, for responding to "*ball present*" (second panel; identical to *KTE's* paradigm), responding to "*ball absent*" (third panel; tested with our Studies 2–3), and responding to ball present in "*occluder*" trials, where there is an occluder between the agent and the ball at all times (fourth panel; tested with our Study 4).

Stimuli. Both groups independently created their own sets of stimuli using the description in KTE's supplementary online materials.[2] The stimuli for the in-lab studies conducted at MIT were, to the best of our knowledge, almost identical to those of KTE, while the stimuli for the online studies were slightly longer but contained essentially the same timing characteristics. Screenshots from the stimuli used for the online studies, as well as the important timings for each stimulus type, are listed in Figure 1 (timings for KTE are estimated from watching the single sample experiment video in their online supplementary material).

Exclusion criteria. Trials with inaccurate responses, either during the attention check or the detection response, were dropped. In addition, participants were excluded if they failed to meet

[2] Neither group was able to obtain the original stimuli used by KTE.

90% accuracy on both the attention check and the detection response across all trials. For example, in Study 1a there were 40 trials, each with an attention check response and a "ball present" response (or non-response if ball was absent). Thus there were 80 responses per participant. Trials were dropped if either or both responses for that trial were incorrect (e.g., if a participant missed the attention check but correctly responded to the ball, the trial was still dropped). Participants were excluded if they made fewer than 72 (90%) correct responses across all trials. For studies with no attention check, we maintained the 90% criterion for the detection responses. This relatively highly exclusion rule ensured that the data collected online through Amazon Mechanical Turk was from participants who were paying careful attention to the task, and was thus comparable to data collected in the lab.

Note that we used the same exclusion criteria across all our studies, and this decision resulted in variable exclusion rates, ranging from 30% (in Study 4) to 0% (in all our in-lab studies). We hypothesize that the exclusion rates are related to how engaged participants were in our task. For example, those participants who completed the in-lab versions of the study were probably more motivated to stay engaged throughout the task, while those who completed the online versions of the task may have been less motivated.

Sample sizes. All planned sample sizes were determined prior to data collection. For the in-lab studies, we chose to collect the same number of participants as in KTE (N=24). For our online studies, we planned to collect a sample that was 2.5 times larger than the original (N=60), which would provide an 80% power to reject a detectable effect (Simonsohn, 2013). Hence, for Study 1a, we collected a sample size of N=60. However, since we then used an exclusion criteria to filter out participants with low accuracy, we subsequently decided to increase the sample size to

N=80 to allow for adequate power even after the exclusion rule was applied. This was true except for Study 3, in which we inadvertently collected N=100 instead.

Statistical Approach. For our initial replication study (Studies 1a–1c), we followed the statistical approach used by KTE, in which separate *t*-tests were used to make pairwise comparisons between conditions. In addition to the issues of multiple comparisons, however, this method does not allow us to characterize the overall pattern of results we obtained. In particular, we observed a highly-consistent and characteristic crossover interaction pattern. This interaction did not conform to the pattern of data previously reported by KTE and it is clearly not predicted by their theoretical account. We discuss this crossover interaction at length below.

Because the crossover interaction describes the relationship between four measurements, it cannot be appropriately tested via independent *t*-tests. Thus, in addition to providing *t*-tests for our initial replication studies, we aggregate information across our follow-up studies by quantifying this crossover interaction. We adopt the following summary approach: We fit a linear mixed effects model (Gelman & Hill, 2007; Jaeger, 2008) using the lme4 package in R (Bates, Maechler, & Bolker, 2012) with the structure `rt ~ participant.belief * agent.belief + (participant.belief * agent.belief | subject)` (this model uses a "maximal" random effect structure; Barr, Levy, Scheepers & Tily, 2013). We then use the reliability of the crossover interaction effect as a test of having observed the same crossover pattern we did in the initial replication studies. When regression coefficients are presented, they are accompanied by their 95% confidence intervals. Because of the large amount of data we collected, we assess normally distributed confidence intervals on coefficients using the $t = z$ method (Barr et al., 2013). Full models, data, and analysis code can be found at

`https://github.com/langcog/KTE`. For statistical results, we report effect sizes – either Cohen's *d* for pairwise comparisons, or regression coefficient *b*'s and 95% confidence intervals in square brackets. We provide units where appropriate, and additionally report *p*-values.

Experiment-specific Methods:

Study 1a (N=60; 6 excluded) and **Study 1b** (N=80; 8 excluded) were conducted online using Amazon Mechanical Turk (all subsequent online experiments were also conducted using the same population). **Study 1c** (N=24; 0 excluded) was conducted in the lab: participants were adults (17 females, ages 18–26, mean 20.6) tested in quiet, dark rooms at the Massachusetts Institute of Technology (MIT). Studies 1a–1c were all direct replications of KTE's Experiment 1.

Study 2a (N=80; 18 excluded) was conducted online. **Study 2b** (N=24; 0 excluded) was conducted in the lab: participants were adults (21 females, ages 18–55, mean 22.58) tested in quiet, dark rooms at MIT. These two studies differed from Studies 1a–1c in that participants were asked to press one button if the ball was present, and to press another button if the ball was absent (i.e. a two alternative forced-choice, or 2AFC, response). This manipulation allowed us to measure responses to both ball-present, and ball-absent, trials. If response times reflect automatic ToM such that participants are automatically encoding the contents of the agent's beliefs, then participants' response to the *absence* of the ball should be facilitated on trials when the agent believes the ball is absent (just as participants' responses to the presence of the ball are facilitated on the trials where the agent believes the ball is present).

Study 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, it required exactly the opposite response criteria from Studies 1a–1c, and unlike Study 2a–2b did not require participants to respond on every trial. This provides an even more minimal pair to Study 1a–1c.

Study 4 (N=80, 23 excluded) was conducted online. The only change from Studies 1a–1b was that an additional permanent occluder was added to entirely obstruct the agent's view of the ball. This study tested whether response times are sensitive to what the agent could see: if participants were implicitly tracking the agent's beliefs, then the presence of this occluder, which obstructed the agent's view, should eliminate effects due to the agent's beliefs.

Study 1–4 Results

We replicate the critical statistical results of KTE (2010). We replicated the main statistical comparisons reported by KTE. KTE's *t*-tests (p. 1832) are reported along with the equivalent tests for Studies 1a–1c in Table 1. 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, compared to when neither did (P+A+ < P-A-; Cohen's d = 0.284, 0.393, 0.654 respectively for Studies 1a,b,c; Cohen's $d = D / s$ where D is the difference between means, and s the standard deviation of the differences. p's = 0.04, 0.001, 0.004). Second, participants were also faster when they believed that the ball was present but the agent did not, as compared to when neither they nor the agent believed that it was present (P+A- < P-A-; d's = 0.611, 0.555, 0.792; all p's < 0.001). These first two comparisons confirm the expected result that the

participant's belief has an effect on the participant's reaction time. Specifically, when the participant believes that the ball is present behind the occluder, the participant is faster to detect the ball, as compared to when the participant expects the ball to be absent (and is presumably surprised by the presence of the ball).

Third, and most importantly, we also replicated the critical result that KTE 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), as compared to when neither believed it was present (P-A+ < P-A-; *d*'s = 0.594, 0.473, 0.422; $p < 0.001$, $p < 0.001$, $p = 0.05$ respectively). Based on this comparison, KTE proposed that the agent's belief facilitated participants' detection of the ball.

Fourth, we replicated the null result that participants' reaction times did not differ between the case when only the agent believed that the ball was present, and the case when only the participant had a belief that the ball was present (P-A+ ~ P+A-; *d*'s = 0.079, 0.062, 0.324; *p*'s = 0.57, 0.60, 0.13).. KTE suggest that both types of beliefs (participant beliefs and agent beliefs) individually facilitate reaction times to the same degree. All the statistical tests that were reported by KTE were replicated in all three studies; this robustness indicates that the effects that KTE reported are highly replicable across different sets of stimuli and different testing environments (online vs. in lab).

Comparison	Study	*t*-statistic	*df*	***p* value**	Cohen's *d*
(P-A-) - (P+A+)	KTE	3.47	23	**0.002**	0.708
	1a	2.09	53	**0.042**	0.284
	1b	3.33	71	**0.001**	0.393

	1c	3.21	23	**0.004**	0.654
(P-A-) - (P+A-)	KTE	3.43	23	**0.002**	0.700
	1a	4.49	53	**<.001**	0.611
	1b	4.71	71	**<.001**	0.555
	1c	3.88	23	**<.001**	0.792
(P-A-) - (P-A+)	KTE	2.42	23	**0.02**	0.494
	1a	4.37	53	**<.001**	0.594
	1b	4.01	71	**<.001**	0.473
	1c	2.07	23	**0.05**	0.422
(P-A+) - (P+A-)	KTE	0.99	23	**0.33 n.s.**	0.202
	1a	0.58	53	**0.57 n.s.**	0.079
	1b	0.53	71	**0.60 n.s.**	0.062
	1c	1.59	23	**0.13 n.s.**	0.324

Table 1. Direct replication of results of Experiment 1 from KTE (2010) using Studies 1a–1c. The *t*, *df*, and *p* values from KTE were reported in the paper, while Cohen's *d* for KTE's studies were calculated from the *t* and *df* values.

We observe a crossover interaction that is not consistent with KTE's theory. In addition to replicating KTE's reported results, we also observed a consistent pattern in the reaction times that KTE did not report (Studies 1a–1c, top row of Fig. 3): all three experiments showed a strong crossover interaction.[3] The interaction coefficients for Studies 1a–1c (with 95% CIs) are: 175 [97, 253], 121 [65, 176], 66 [18, 114] msec ($p < 0.001$, $p < 0.001$, $p = 0.007$ respectively). The

[3] For our online studies, reaction time was recorded from the first frame in which the occluder started to be lowered; the occluder took 200 msec to fall completely. From the details of KTE's report it was unclear when timing began and hence we cannot directly compare mean reaction times between our experiments and those of KTE, as there may be constant differences between experiments in when timing began. The critical question concerns the pattern of reaction times between conditions.

crossover was caused by relatively slow reaction times on P+A+ trials. If reaction times 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+ *slower* than P+A-; *d*'s = 0.35, 0.20, 0.41; *p*'s = 0.01, 0.09, 0.06). This crossover interaction is thus not consistent with automatic ToM, and it was not observed in the data that KTE report (Fig. 1)[4]. Nevertheless, this crossover interaction was robustly present in all three of our replications (as well as in our subsequent studies, reported below). Hence, although we consistently replicated all of KTE's reported statistical tests, 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 the interpretation of this pattern in terms of automatic ToM comes from Studies 2a–2b and 3. Recall that the prediction based on a ToM account is 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 previous pattern should no longer be observed). In Study 2a and 2b, participants responded to both ball presence and ball absence. The trials of interest are the correct rejections ("CR"), where participants correctly indicate that the ball is absent. In Study 3, participants only responded to the absence of the ball; the results of these studies are shown in Fig. 3.

If reaction times reflect automatic ToM, participants should be faster (or at least not slower) to respond to the *absence* of the ball when the agent correctly believed the ball was

[4] Without access to KTE's original data we cannot make a direct test of whether our results differ reliably from KTE's. Given their relatively small sample size and large confidence intervals, however, it is possible that – although there was no crossover observed in their data – their results and ours are not inconsistent.

absent (P-A-) than when the agent falsely believed the ball was present (P-A+), as illustrated in Fig. 2 (top row, right panel). Contrary to this prediction, participants were *faster* to respond to the ball's absence for P-A+ than for P-A- (P-A+ faster than P-A-; *d*'s = 0.42, 0.81, 0.66 for Study 2a CR, 2b CR, 3 respectively; all *p*'s < 0.001). Moreover, we observed exactly the same crossover pattern of reaction times across conditions for responses to the ball's absence as we did for responded to the ball's presence (Fig. 3, interaction *b* in Study 2a CR, 2b CR and 3 are: 207 [114, 301], 161 [102, 221], 173 [116, 229] msec; all *p*'s < 0.001).

We next collapsed across Studies 1–3 and tested whether the CR and "absent" trials produced a different pattern of reaction times as the original "present" trials. The model for this analysis included terms for participant and agent belief, as well as a term for whether the trials were CR/absent trials (and all interactions). Although we found a main effect of CR/absent trials, which were overall slightly slower (*b* = 68 [39, 97] msec; *p* < .0001), there were no reliable two- or three-way interactions with CR/absent trials (all *b*'s < 44 msec, all *p*'s > .10). In addition, the two-way interaction between participant and agent beliefs that we observed in each individual experiment was still reliable (*b* = 139 [105, 172], *p* < .0001). This analysis thus supports the claim that, across studies, there was no statistical difference in the pattern of reaction times 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' reaction times reflect automatic encoding of the agent's belief, we replicated Study 1 with one critical difference in the stimuli: a substantially large wall blocked the agent's view. In this study, the agent has no perceptual access to the ball; thus the response time should be affected only by the participants' own belief (compare the theoretical prediction

in Fig. 2, last panel, with the data shown in Fig. 3, bottom right panel). Yet, contra that prediction, the pattern of reaction times across conditions remained similar to previous experiments, and the crossover interaction was still reliable (interaction b = 109 [49, 169] msec; $p < 0.001$).

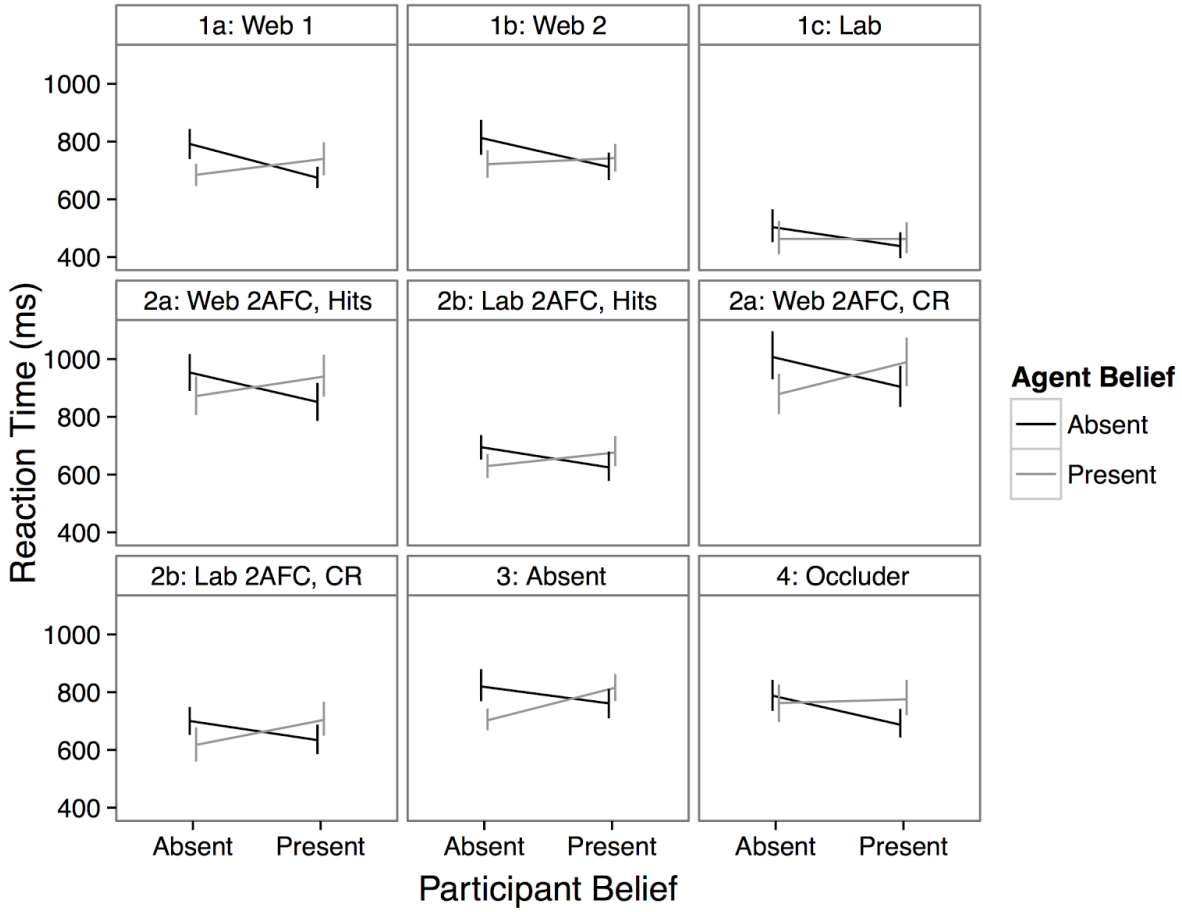

Figure 3: Mean reaction times by condition and experiment. The crossover interaction is statistically reliable for every experiment and condition (see text for interpretation). Error bars represent 95% confidence intervals of the mean. Lines are displaced slightly along the horizontal axis for clarity. Top row from left to right: Studies 1a, 1b, and 1c (direct replications). Middle row: Study 2a, 2b Hits (respond "present" *when* ball is present), and 2a Correct Rejections (CRs; respond "absent" when ball is absent). Bottom row**:** 2b Correct Rejections, Study 3 (respond "absent" when ball is absent), and Study 4 (permanent occluder between agent and ball).

Discussion

We replicated KTE's critical results in three studies, using two stimulus sets, with equal or substantially greater power (Studies 1a–1c). By varying when participants respond (i.e. altering whether they respond to 'present' vs 'absent' trials), and by varying the agent's beliefs (i.e. eliminating the agent's perceptual access to the ball), Studies 2–4 tested whether differences in reaction times 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 reaction times across conditions in Studies 1–4 was better characterized by a cross-over interaction than by two main effects (see Fig. 2). Importantly, this pattern of reaction times (i) is not predicted by automatic belief tracking and (ii) seems to be driven by some difference in the stimuli across conditions that is *independent* of the relationship between the agent's belief and the ball's final position. Because of the robustness of the crossover interaction, we set the primary goal of Studies 5–8 as understanding 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). While the attention check was a minor point in KTE's experimental design and was only mentioned in their Supplementary Online Materials, its timing varied substantially across conditions. In order to allow for the agent's beliefs to differ in the different condition, the original stimuli manipulated the time at which the agent left the scene, which coincidentally also varied the time at which participants were required to press an additional button to indicate that they were paying attention. This suggests that the timing of the attention check may have generated the pattern of reaction times in KTE's paradigm. One piece of supportive evidence for this hypothesis comes from Kovács, Kühn, Gergely, Csibra, and Brass (2014), who removed the

attention check and used a 2-alternative, forced-choice paradigm as in our Experiment 2. Consistent with the hypothesis that the attention check was responsible for the pattern of reaction times in KTE, with these modifications, they observed no reliable reaction time differences between conditions.

We hypothesized a specific mechanism by which the attention check confound could have generated the pattern of results we observed. For two sequential judgments, the shorter the period of time between them, the slower the second judgment tends to be. This finding is sometimes known as the "psychological refractory period." Intuitively, having just made a quick response, it is harder to immediately make another second quick response. A large literature exists attesting to this finding and investigating its underlying mechanisms (e.g., Telford, 1931; reviewed in Herman & Kantowitz, 1970). In the case of KTE's paradigm, when the attention check occurs later in the trial, participants' response times could be slowed by the relative shortness of the "refractory period" before they are required to indicate the presence or absence of the ball. In KTE's original design, shorter delays between their attention check and the primary response were confounded with belief condition. These variable delays may thus have generated the observed pattern of results.

We test this "attention check" hypothesis in the next set of studies. Studies 5–6 demonstrate that the timing of the attention check plays a critical role in affecting participants' reaction times. Studies 7–8 go on to provide a double dissociation: the crossover effect demonstrated in Studies 1–4 is present when the attention check is present (regardless of the agent's presence or beliefs), but absent when the attention check timing is controlled (again, regardless of the agent).

Study 5–8 Methods

General Methods. The methods used in Studies 5–8 are identical to those in Studies 1–4 except where explicitly noted. Similar to Studies 1–4, participants were shown videos of a ball and an occluder on a table. Each trial consisted of watching a brief video and making a judgment about the ball displayed in the video. Each study required participants to complete 40 trials (8 movies 5 times) except where indicated otherwise. Participants were recruited through Amazon Mechanical Turk, were self-paced, and took an average of 20–25 minutes. Informed consent was obtained on the first page of the experiment, and Studies 5–8 were all approved by the Stanford University Institutional Review Board. These studies were collaboratively conducted by all authors.

Exclusion Criteria. For consistency across studies, we applied the same 90% accuracy exclusion rule to Studies 5–8. We again note that this decision resulted in variable exclusion rates, ranging from 30% (in Study 7) to 1–2% (in Studies 6 and 8a). Such differences are to be expected because of the variation in how engaging the tasks were. For example, Study 7 had no agent but contained an unpredictably-timed light bulb; this experiment was likely both boring and difficult for participants. In contrast, Study 6 contained an agent and had a predictably-timed attention check; it was hence probably both easier and more engaging.

Sample Sizes. Because of the exclusion criteria used to filter out participants with low accuracy, we chose a sample size of N=80 to allow for adequate power (i.e, 80% power to reject a detectable effect), even after exclusion (Simonsohn, 2013). This was true except for Studies 5b

and 8b, in which we increased the sample size to 200 as an *a priori* decision 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 effect model to capture the size of the crossover interaction in a way that is relatively comparable across studies, with all details identical to those presented in the previous studies.

Experiment Specific Methods

Study 5a (N=80; 16 excluded) was conducted online. The only difference from Studies 1a–b was that the attention check was removed, i.e., participants did not have to respond when the agent left the scene.

Study 5b (N=200; 25 excluded) was conducted online and contained a replication of Study 5a and Study 1, designed for matched statistical comparison. We planned a sample of 200 to ensure adequate power to test for a three-way interaction. In this experiment, participants completed two blocks. One block was a replication of Study 1, i.e. with the attention check, and the other block was a replication of Study 5a, i.e., without the attention check. The block order was counterbalanced across participants. Within each block, participants viewed each of the 8 videos 3 times; thus, there were 24 trials per block and hence, an increased number of trials of 48.

Study 6 (N=80; 1 excluded) was conducted online. The only difference from Studies 1a–1b was that the attention check was instead moved to when the agent returned rather than when he left,

because this event occurred at the same time (19s) in all of the conditions. Thus, this study held the timing of the attention check constant while still ensuring participant compliance.

Study 7 (N=80; 24 excluded) was conducted online. In this study, the agent was removed and replaced with a stationary light bulb that flashed on at the time when the agent would have left the scene. The light bulb flashed on once and then remained on through the duration of the trial. The light bulb's flashing on was then used as a modified attention check that 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.

Study 8a (N=80; 2 excluded) was conducted online. In this study, the agent was present (as in Studies 1–6), and there was also a light bulb present. As in Study 7, the light bulb's flash was used as an attention check. However, in contrast to Study 7, the time at which the light bulb flashed on was completely dissociated from the agent's and participant's beliefs. Thus, the 3 possible attention check timings that were present in the original videos {10.8s, 13.2s and 16.7s} were crossed with the 4 belief conditions {P+A+, P+A-, P-A+, P-A-} and the presence of the ball {present, absent}, leading to 3*4*2 = 24 possible videos. Each participant viewed each of the 24 videos twice, which resulted in an increase of the total number of trials to 48.

Study 8b (N=200; 37 excluded) was conducted online. It was identical to Study 8a, except that we tested a set of five evenly spaced attention check timings {10.9s, 12.9s, 14.9s, 16.9s and 18.9s}, fully crossed with the 8 possible videos. The attention check timings were designed to be 2s apart, and were chosen to span the range of attention check timings we had tested earlier, from

the minimum of 10.8s to the maximum of 19s. The 5 timings x 4 belief conditions x 2 for the presence or absence of the ball resulted in 40 different videos. To maintain the experiment length, each participant viewed each video only once, for a total of 40 trials. We predicted that having only one repetition per video would result in more noise, and hence we chose to increase our sample size to add additional statistical power.

Study 5–8 Results

The crossover interaction is observed only when there is an "attention check" with variable timing. In Study 5a, the attention check requirement was removed, and the response time pattern became flat, with no crossover interaction (interaction b = 22 [-47, 91] msec; p = 0.53).

Study 5b provided an additional replication of Study 1 and Study 5a in a within-subjects design, to allow for direct statistical comparisons between the two. In two blocks of 24 trials (6 trials per condition, with blocks in a random order across participants), participants were either asked to respond to the attention check or not. In this study, we found a reliable three-way interaction of participant condition, agent condition, and attention check condition (three-way interaction b = 76 [8, 145] msec, p = 0.029). There was still a crossover-interaction even when there was no attention check (interaction b = 62 [-16, 140] msec, p = 0.036), but the size of the effect was more than doubled in trials where there was an attention check (b = 140 [88, 192] msec, p < .001). The three-way interaction provides evidence that the magnitude of the crossover observed in Study 1, but not in Study 5a, is driven by the attention check.

Summarizing these results, Studies 5a and 5b show that removing the attention check reduces differences in RT across conditions. However, this experiment does 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, keeping them from encoding either participant or agent beliefs.

To address this issue, in Study 6, the attention check was shifted to when the agent returned to the scene, which was at 19s in all conditions.[5] Once again, the pattern of responses was flat (interaction b = 12 [-35, 59] msec; p = 0.62). This study used the exact same stimuli as Studies 1–3, except that the attention check timing was matched across all four conditions, again based on a salient action of the agent. Critically, the characteristic pattern of response times found in Studies 1–3 was absent.[6]

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

The pattern of observed reaction times is a parametric function of the timing of the attention check and is independent of belief condition and even the presence of the agent. 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 in that condition. To make this possible, we included a light bulb in the videos and instructed participants to press an additional button when the light bulb came on. This event was then used as the attention check

[5] The agent returned 2s before the occluder was lowered, so we reduced the "attention check window" from 3s to 2s in this study.

[6] The unusually fast overall response times in Study 6 compared to the other studies most likely arose because, in this study, the attention check reliably appeared 2 seconds before the occluder fell in all conditions, preparing participants to respond exactly 2s afterwards. It should not be inferred that the later attention checks facilitate detection more generally, but rather that the *predictability* of the attention check facilitated subsequent responding. In fact, this predictability effect is found routinely within the literature on simple reaction time studies cited above and provides further evidence for our contention that RT measurements in KTE's paradigm are extremely sensitive to features of the attention check (both its relative timing and its predictability).

instead of the agent's departure. (As before, all other aspects of the studies remained identical, except where noted). By replicating the asymmetric attention check pattern in the absence of an agent (Study 7), and by varying the attention check independent of the agent (Study 8), we were able to test for a complete dissociation between attention check timing and belief condition.

In Study 7, we removed the agent entirely but had the light bulb differentially switch on at the times that corresponded to when the agent left the scene in Studies 1–4 (i.e., 10.8s, 13.2s and 16.7s, see Fig. 1.). As in Studies 1–4, participants were instructed to press an additional button to indicate that they had been paying attention. Thus, participants were asked to respond at the exact same times in Study 7 as they were in Studies 1–4. We once again observed a crossover interaction (interaction b = 86 [32, 140] msec; p = 0.002), though it was slightly smaller than before. This time, however, the crossover interaction was observed without an agent being present at all! Thus, the results of Study 7 support the hypothesis that the response times observed in Studies 1–3 were independent of agent beliefs, and were plausibly driven by the attention check.

Study 7 showed that the reaction time difference between conditions can be elicited without an agent but with the corresponding attention check timing. Study 8 goes further by showing that, even when the agent is present, the reaction time effect remains absent if the attention check timing is appropriately controlled.

Study 8a crossed the timing of the light bulb flash with the video condition: 3 timings (10.8s, 13.2s, 16.7s) crossed with 8 belief condition videos. Study 8b used 5 evenly spaced timings when the light bulb switch on (10.9s, 12.9s, 14.9s, 16.9s and 18.9s), again crossed with the 8 videos.[7] As in Study 7, the participant was instructed to press a button when the light bulb

[7] Note that there were only two trials per condition in Study 8a and one trial per condition in Study 8b.

flashed. Averaging across attention check timings, there was no cross-over interaction in RTs based on belief condition in either study (Fig. 4; Study 8a: interaction b = 5.3 [-38, 48] msec; p = 0.81; Study 8b: interaction b = -32.6 [-67, 2]; p = 0.07).

To test the effect of attention check timing on subsequent ball-detection RT, controlling for belief condition, we added attention check timing as a continuous predictor variable in our regression model (which, as discussed above, fits separate coefficients for participant and agent beliefs and their interaction). This model showed a reliable linear effect of attention check timing in both studies (coefficient on the attention time = 9.7 [5.5, 13.9] and 12.1 [9.1, 15.1] msec/sec; p's < 0.001, Fig. 5). The closer to the ball-detection decision the attention check was, the slower the ball-detection decision was. As discussed above, this result is congruent with literature on the psychological refractory period, which suggests that the offset between two reaction-time measurements has systematic effects on the latency of the second measurement.

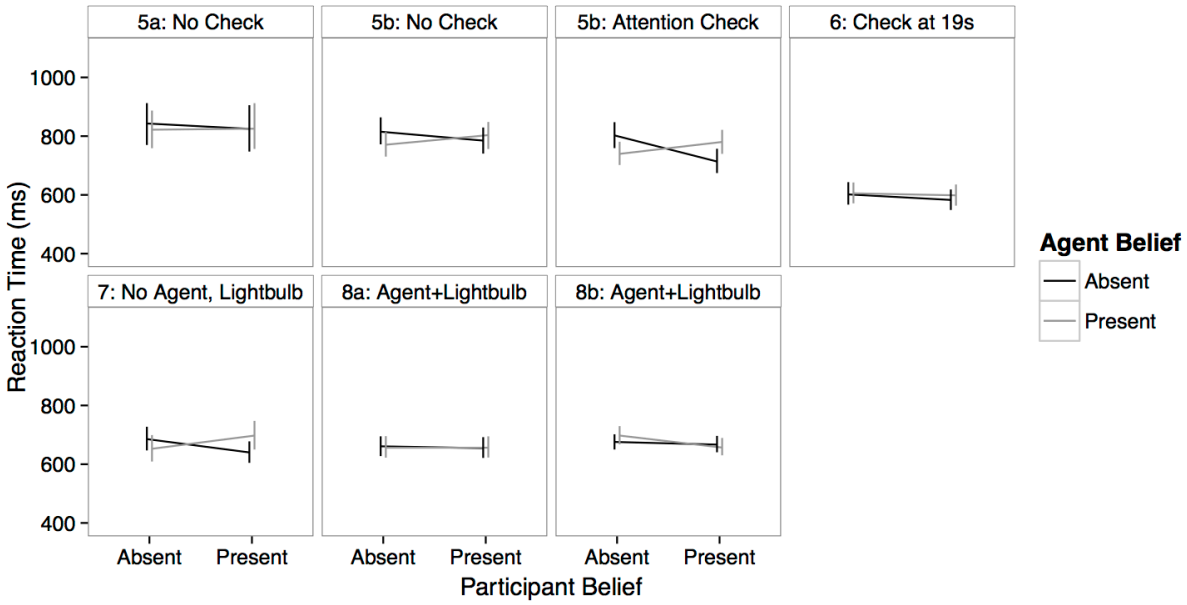

Figure 4: Reaction times by condition and experiment. Crossover interaction was only statistically reliable in Study 5b and Study 7 (see text). Error bars represent 95% confidence intervals. Lines are displaced slightly along the horizontal axis for clarity. Top row, from left to

right: Study 5a (attention check was removed), Study 5b trials with attention check removed, Study 5b trials with attention checks, and Study 6 (attention check was moved to the same time for all videos). Bottom row: Study 7 (agent was removed, and participants had to respond to the flash of a light bulb as an attention check), and Study 8a and 8b (agent is present, but participants responded to the flash of a light bulb at different times).

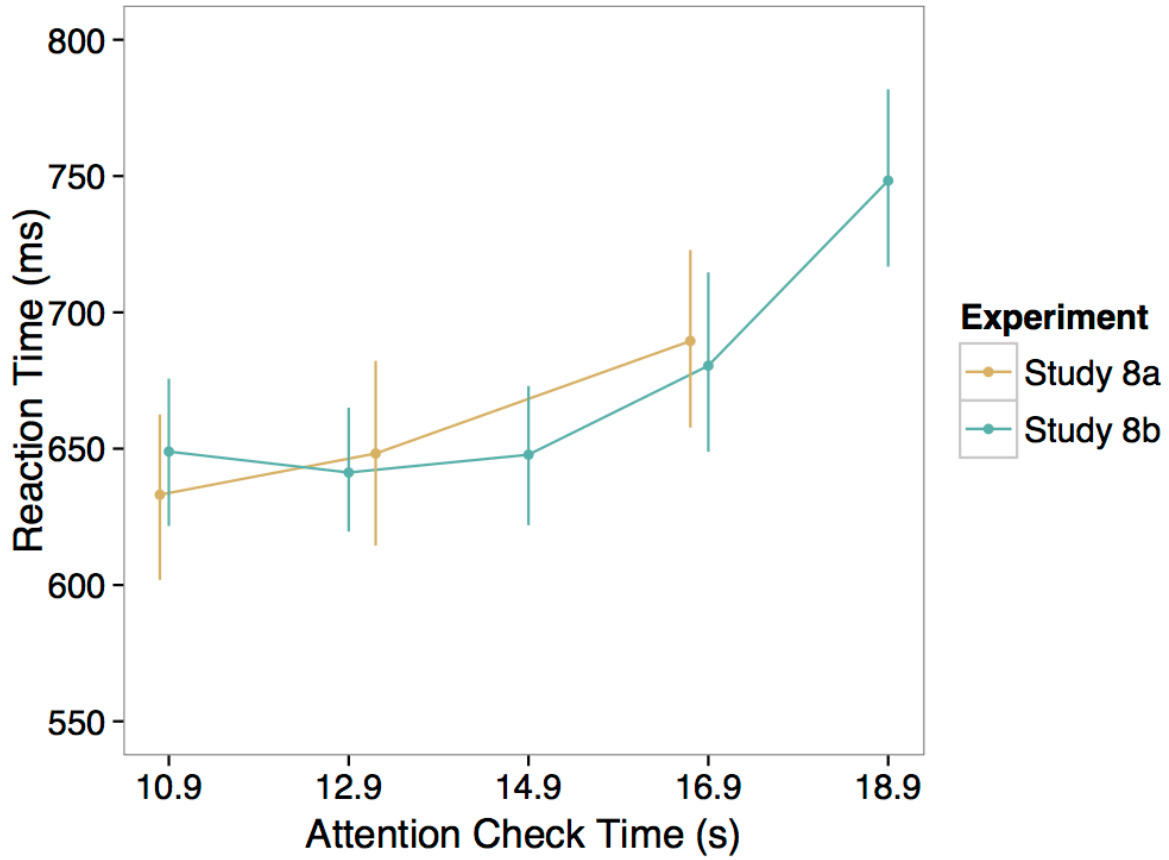

Figure 5: Studies 8a and 8b, decomposed by the attention check timing and demonstrating slower response times with a progressively later attention check. Error bars represent 95% confidence intervals.

General Discussion

Collectively, these eight studies provide good reason to reconsider the primary evidence for automatic ToM in human adults. We began by robustly replicating KTE's key effects that they claimed supported the automaticity of ToM in adults (Studies 1a–1c). Studies 2–4 then demonstrated that these effects are not sensitive to the content of the agent's beliefs or

perspective. In Studies 5–8, we test the hypothesis that KTE's key effects may have instead arisen from an artifact of the specific method used. Studies 5 and 6 demonstrate that the effect is related to the presence and timing of the attention check in the paradigm used to ensure participant compliance. Studies 7–8 go on to show a double dissociation. The reaction time effect can be replicated without any agent at all, provided that the attention check timing is asymmetric across conditions (Study 7), and the reaction time effect is completely absent even when the agent is present, provided that the attention check timing is not confounded with the agent's beliefs (Studies 8a–8b). In sum, the evidence across all these studies is inconsistent with an automatic ToM account.

These studies also provide one plausible mechanistic account of the attention check timing artifact: The two conditions whose reaction times were consistently highest in our replications (P+A+ and P-A-) also showed the shortest delay between the attention check task and the primary ball-detection response. The quick succession of the cue and the primary response likely led to increased reaction times via the same "refractory" mechanism at play in completely non-social reaction-time tasks (Telford, 1931; Herman & Kantowitz, 1970).

Stepping back to consider Studies 1–8 as a whole, we conducted two random-effect meta-analyses. Figure 6a summarizes the magnitude of the crossover interaction effect across conditions, grouped by the predictions of the attention check hypothesis. Figure 6b provides a meta-analytic summary of the P-A+ > P-A- comparison (the critical test for automatic ToM provided by KTE), grouped by the predictions of the automatic ToM hypothesis. The attention check hypothesis does an excellent job of accurately explaining the results of our studies, 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.

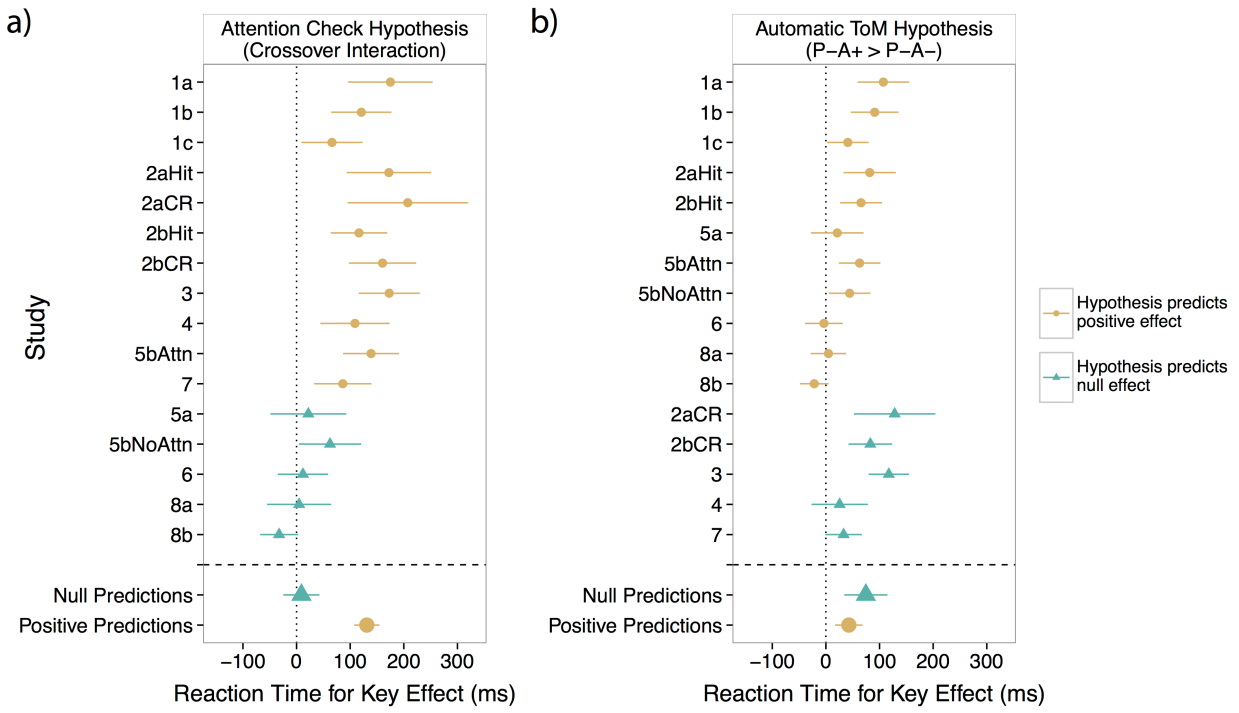

Figure 6. **(a):** Meta-analysis of the magnitude of the effect sizes of the crossover interaction for all the studies reported in this paper. The magnitudes are plotted along the horizontal axis as circles, and error bars represent 95% CIs. A solid vertical line at 0 msec is overlaid for reference. The effect sizes are grouped by whether the studies are predicted by the "Attention Check Hypothesis" to show a crossover interaction (i.e. whether or not the belief conditions differed by attention check timing). The bottom two points show the meta-analytic effect size for the "null prediction" and "positive prediction" studies, calculated using a random-effect meta-analysis (Borenstein, Hedges, & Higgins, & Rothstein, 2010). **(b):** A similar meta-analysis, but performed for the critical test of automatic false belief representation in KTE (2010): the difference in reaction times between P-A+ and P-A-. Note that effects are ordered differently between the two panels.

In conclusion, while KTE's (2010) results are highly replicable, they do not provide evidence for automatic belief computation in human adults. The related evidence KTE provide for ToM in preverbal infants is, by contrast, not undermined by the studies we present. Yet at the same time, the present work does clearly demonstrate that the stimuli employed used in those studies 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 studies do not provide conclusive evidence *against* automatic ToM. Rather, the present work highlights the need for new investigations into this aspect of human cognition. We are currently only aware of 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.

References

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

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

Baron-Cohen, S. (1989). The Autistic Child's Theory of Mind: A Case of Specific Developmental Delay. *Journal of Child Psychology and Psychiatry, 30*(2), 285-297. doi:10.1111/j.1469-7610.1989.tb00241.x

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(3), 255-278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M. & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4

Bennett, J. (1978, 01). Some remarks about concepts. *Behavioral and Brain Sciences, 1*(04), 557. 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(2), 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*(2), 73-92. doi:10.1002/icd.298

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

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(04), 568. doi:10.1017/S0140525X00076664

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives on Psychological Science, 7*(6), 600-604. doi:10.1177/1745691612460686

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. 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*(3), 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*(3), 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*(1), 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(4), 434-446.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 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*(1), 113-122. doi:10.1111/j.1467-7687.2011.01098.x

Kovács, A. M., Teglas, E., & Endress, A. D. (2010). The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults. *Science, 330*(6012), 1830-1834. doi:10.1126/science.1190792

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), e106558. doi:10.1371/journal.pone.0106558

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*(3), 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*(2), 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*(5719), 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*(6), 657-660. doi:10.1177/1745691612462588

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *Behavioral and Brain Sciences, 1*(04), 592. doi:10.1017/S0140525X00076895

Simonsohn, U. (2013). Small Telescopes: Detectability and the Evaluation of Replication Results. Working paper (dated Dec 10, 2013), available at SSRN: http://ssrn.com/abstract=2259879.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of Beliefs by 13-Month-Old Infants. *Psychological Science, 18*(7), 580-586. doi:10.1111/j.1467-9280.2007.01943.x

Surian, L., & Geraci, A. (2011). 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), 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*(1), 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*(1), 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*(1), 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, 104*(20), 8235-8240. doi:10.1073/pnas.0701408104