Psychological Research
https://doi.org/10.1007/s00426-019-01181-7

ORIGINAL ARTICLE

Measuring spontaneous mentalizing with a ball detection task: putting the *attention-check hypothesis* by Phillips and colleagues (2015) to the test

Rachida El Kaddouri[1] · Lara Bardi[2,3] · Diana De Bremaeker[1] · Marcel Brass[2] · Jan R. Wiersema[1]

Received: 26 October 2018 / Accepted: 29 March 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Theory of Mind (ToM) or mentalizing refers to the ability to attribute mental states (such as desires, beliefs or intentions) to oneself or others. ToM has been argued to operate in an explicit and an implicit or a spontaneous way. In their influential paper, Kovács et al. (Science 330:1830–1834, 2010) introduced an adapted false belief task—a ball detection task—for the measurement of spontaneous ToM. Since then, several studies have successfully used versions of this paradigm to investigate spontaneous ToM. This paradigm has, however, been criticized by Phillips et al. (Psychol Sci 26(9):1353–1367, 2015), who argue that the effects are fully explained by timing artifacts in the paradigm, namely differences in timing of the attention check. The main objective of the current study is to test this *attention-check hypothesis.* An additional aim was to relate the findings to autism spectrum disorder (ASD) symptomatology in our neurotypical sample, as ASD has been linked to deficits in spontaneous mentalizing. We applied an adjusted version of the paradigm in which the timings for all conditions are equalized, ruling out any potential timing confounds. We found significant main effects of own and agent beliefs on reaction times. Additionally, we found a significant 'ToM-effect': When participants believe the ball is absent, they detect the ball faster if the agent believes the ball would be present rather than absent, which refers to the original effect in the paper of Kovács et al. (2010), taken as evidence for spontaneous ToM and which was contested by Phillips et al. (2015). Our findings cannot be explained by the attention-check hypothesis. Effects could not be associated with ASD symptoms in our neurotypical sample, warranting further investigation on the link between spontaneous mentalizing and ASD.

Introduction

Social interactions are driven by the ability to attribute mental states (beliefs, intentions, desires and feelings) to oneself and others, which is referred to as Theory of Mind (ToM) or mentalizing (Premack & Woodruff, 1978). ToM has been argued to operate in either an explicit or implicit/spontaneous mode. Explicit mentalizing refers to a cognitive process during which a person is deliberately considering mental states of others (Schuwerk, Vuori, & Sodian, 2015; Wellman, Cross, & Watson, 2001). Implicit or spontaneous mentalizing is delineated as a rapid, inflexible, cognitive efficient process that operates without being consciously aware of it (Clements & Perner, 1994; Kulke, von Duhn, Schneider, & Rakoczy, 2018; Nijhof, Brass, Bardi, & Wiersema, 2016; Schneider, Slaughter, & Dux, 2017; Schuwerk et al., 2015). Past research has mainly focused on explicit mentalizing as the notion of spontaneous mentalizing has only arisen recently, but research on spontaneous mentalizing is rapidly expanding. Investigating spontaneous mentalizing may not only provide a better understanding of development of ToM (Low & Perner, 2012; Schneider, Slaughter, Becker, & Dux, 2014; Schneider et al., 2017) but may also be of major relevance for studying ToM in psychopathology, such as autism spectrum disorder (ASD; Senju, 2013a, b). ASD is a neurodevelopmental disorder characterized by qualitative

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00426-019-01181-7) contains supplementary material, which is available to authorized users.

✉ Rachida El Kaddouri
Rachida.ElKaddouri@UGent.be

[1] Department of Experimental Clinical and Health Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

[2] Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

[3] Institute of Cognitive Sciences Marc Jeannerod, UMR 5229, CNRS and University of Lyon, Lyon, France

Published online: 11 April 2019

impairments in social interactions and communication in daily life, which has been explained by a mentalizing deficit: People with ASD have difficulties understanding other people's mental states (Hill & Frith, 2003; Sabbagh, 2004; Schneider, Slaughter, Bayliss, & Dux, 2013). However, findings from studies using explicit ToM tasks are not conclusive, as children and adults with ASD often pass such tests. It has therefore been reasoned that individuals with ASD do not mentalize spontaneously, but may succeed on ToM tasks in which they are explicitly asked about the others' mental state by means of learned cognitive compensatory strategies (Frith, 2012). A deficit in spontaneous ToM may explain why individuals with ASD keep showing severe mentalizing difficulties in daily social life, which is complex and requires fast online implicit mentalizing abilities (Frith, 2012).

To measure spontaneous mentalizing, a variety of paradigms has been developed to allow investigating mental state attribution without requiring participants to explicitly deliberate about other people's mental states (Apperly & Butterfill, 2009). Kovács, Téglás and Endress (2010) introduced an adapted false belief task, in which the participant and agent form a belief about the presence of a ball, followed by an outcome phase in which the ball is either present or absent. More specifically, participants get to see short movies in which an agent (in their version: a Smurf) forms a belief about the location of a ball. The ball can either be behind an occluder or roll out of the scene. The agent walks out of the scene, and while he is away the participant forms a belief about the ball's location as well. At the end, the agent walks back in and the participant has to press a button if he/she thinks the ball is present behind the occluder (in the adaptation for infants, no button press is required since eye tracking is used). However, whether or not the ball is behind the occluder (50% of the trials) is independent of what happens during the movie. Kovács et al. (2010) employed this paradigm in two different samples. First, they ran experiments with neurotypical adults and predicted that they would detect the ball faster when they believed the ball would be behind the occluder. More important though, they hypothesized that although the belief of the agent is completely irrelevant to the task, if participants would spontaneously track the belief of the agent, reaction times should be affected by it. They observed that when participants do not expect the ball to be present (P−), they detect the ball faster when the agent believes the ball is present (P−A+) rather than absent (P−A−), which was taken as evidence for spontaneous mentalizing of the agent belief (Kovács et al., 2010). The difference in reaction times between these two conditions was in later studies referred to as the 'ToM-index' (e.g., Deschrijver, Bardi, Wiersema, & Brass, 2016). In a second series of experiments, 7-month-old infants were tested using a violation of expectation paradigm. Instead of investigating reaction times, looking durations were measured. This was indicated by how long the infants looked at the absence of the ball when the participant and/or agent believed the ball would be behind the occluder. As in adults, infants as young as 7 months old seem to spontaneously track the beliefs of the agent (Kovács et al., 2010).

Since then, this paradigm, with some adjustments, has been applied in several studies using brain imaging and in psychopathological groups (e.g., Bardi, Desmet, Nijhof, Wiersema, & Brass, 2017; Deschrijver, Bardi, Wiersema, & Brass, 2016; Nijhof, Bardi, Brass, & Wiersema, 2018; Nijhof et al., 2016; Nijhof, Brass, & Wiersema, 2017; Phillips et al., 2015). These studies revealed three important insights.

Firstly, in all these studies the ToM-index was found, indicating that the ToM-index is a reliable measure that can be replicated in different laboratories (Deschrijver et al., 2016; Kovács, Kühn, Gergely, Csibra, & Brass, 2014; Kovács et al., 2010; Nijhof et al., 2017; Phillips et al., 2015). Secondly, performance during the implicit version of the task elicits brain activation in core ToM regions, such as the temporo-parietal junction (TPJ; Bardi et al., 2017; Kovács et al., 2014; Nijhof et al., 2018). These results support the validity of the paradigm for measuring mentalizing processes (Bardi et al., 2017; Nijhof et al., 2018). Finally, this paradigm has also been employed in relation to ASD to investigate the hypothesis that spontaneous mentalizing abilities are impaired in people with ASD (Deschrijver et al., 2016; Frith, 2012; Kulke et al., 2018; Schneider et al., 2013). Deschrijver et al. (2016) predicted a smaller ToM-index in adults with ASD. They found a significant ToM-index, but groups did not significantly differ for this effect. However, within the ASD group there was a negative correlation between the ToM-index and ASD symptomatology, suggesting less spontaneous mentalizing in adults with ASD showing more ASD symptoms. Nijhof, Brass and Wiersema (2017) tested neurotypical adults with higher and lower levels of ASD symptomatology based on scores on the short Autism Spectrum Quotient (AQ). As expected, participants with a higher level of ASD symptomatology showed less spontaneous mentalizing as indicated by a smaller ToM-index (Nijhof et al., 2017). Finally, Nijhof et al. (2016) found no correlation between the ToM-index and ASD symptomatology within a neurotypical sample (Nijhof et al., 2016). While these studies provided some evidence for the idea that spontaneous mentalizing is impaired in ASD, the results are not unequivocal and the relation between spontaneous ToM and ASD symptomatology may be more subtle than assumed (Deschrijver et al., 2016; Nijhof et al., 2016, 2017).

While the paradigm has been proven to be a useful tool for studying spontaneous mentalizing processes in these studies, the validity of the paradigm as used in adults has been questioned by Phillips et al. (2015). The authors argue that the initial findings of Kovács and colleagues, taken as evidence for spontaneous ToM, are driven by inconsistencies in the timing of an attention check—*the attention-check hypothesis*. The crucial aspect in the paradigm is that the beliefs formed by the agent are completely irrelevant for the task (detecting the ball), but are hypothesized to influence the reaction times as the participant spontaneously takes into account the belief of the agent. To ensure that participants pay attention to the video and the agent, without providing a rational for the presence of the agent, participants are instructed to press a button when the agent leaves the scene (the so-called 'attention check'). Phillips et al. (2015) claim that the difference between conditions in the timing of the attention check is what explains the results (the so-called *attention-check hypothesis).* According to the authors, this is due to the *psychological refractory period* (PRP): The shorter the time between two judgments is, the slower one is on the second judgment (Phillips et al., 2015).

The attention-check hypothesis has been contested by Nijhof et al. (2017). They suggested several theoretical and statistical arguments why this explanation is unlikely. Among others, they did not observe a crossover effect in their data which would be expected as argued by Phillips et al. (2015) based on the timing differences in attention check between conditions. They further criticized the proposed underlying mechanism, the PRP, as the PRP has been known to only have a short-term effect lasting up to several hundred milliseconds, while the shortest interval between the attention check and ball detection was more than 3 s, which exceeds the reach of a PRP effect (Nijhof et al., 2016). Finally, they argued that a negative correlation between the ToM-index and ASD symptom severity in adults with ASD, as found by Deschrijver et al. (2016), is difficult to reconcile with a simple timing explanation. We would like to add to these arguments that neuroimaging findings from studies applying this paradigm showed activation in core mentalizing regions in the brain (e.g., TPJ) and also strongly suggest that mentalizing processes are at play and that these effects cannot easily be explained by timing artifacts in the paradigm (Bardi et al., 2017; Nijhof et al., 2018).

However, none of these arguments form a definitive proof against the attention-check hypothesis. Philips et al. (2015) manipulated the attention check within the paradigm in two different ways to validate their attention-check hypothesis. The results showed that (1) if there is no attention check or (2) if the attention check is at the same time in every condition, namely when the agent comes back into the scene (and not when he leaves the scene, which is the moment the agent forms his belief), the results of Kovács et al. (2010) are not replicated. They therefore concluded that the findings in reaction time patterns are driven by differences in timing of the attention check and not by the belief of an agent. However, their approach is not ideal: It does result in equal timing of the attention check; however, it also may undermine the purpose of the attention check itself. The attention check is there to ensure that the participant pays attention to the agent during a pivotal phase of the movie, namely when the agent forms his belief. Omitting the attention check or postponing it to the final stage of the movie may result in not paying attention to this pivotal phase of the movie. In order to address this concern, we kept the attention check at the same moment as in the original paradigm (when the agent leaves the scene) but ensured equal timing for the attention check between all conditions. In addition, to completely eliminate any timing issue, we ensured equal timing of all events in the paradigm. By equalizing all the timings (how long the agent is in the scene; for how long the ball moves; the moment the agent leaves (thus, the agent has formed a belief and the participant has to press a button); the moment when the agent comes back in; and when the occluder falls), we rule out any explanation in terms of timing confounds. We hypothesized that the reaction time pattern in neurotypical adults will be influenced by both own and agent beliefs. In line with the original findings of Kovács et al. (2010), we expected to find a significant ToM-index. As an additional aim, we wanted to test whether ToM-index scores are related to ASD symptomatology as a dimensional trait in our neurotypical sample. More specifically, we explored whether participants scoring higher on ASD symptomatology show smaller ToM-effects compared to participants with lower scores.

Methods

Participants

Sixty participants (five male; mean age = 18.82 years; SD = 1.41 years) took part in the study. Three participants in total were excluded from further analysis due to an accuracy lower than 90%. Data analysis was thus carried out on data of 57 participants (three male, mean age = 18.82 years; SD = 1.44 years). All of the participants were students at Ghent University and received course credits in return. Informed consent was obtained from all individual participants included in the study. This study was approved by the local ethics committee of the Faculty of Psychology and Educational Sciences of Ghent University.

Stimuli and task

Implicit mentalizing task The task was presented on a laptop (15.6 inch) using Presentation Software, version 18.1 (Neurobehavioral Systems Inc., San Francisco, CA).

An adapted version of the implicit Theory of Mind task (Kovács et al., 2010) was used. The agent in the original version was a Smurf, while the one used in the adapted version was Buzz Lightyear (see also Bardi et al., 2017; Deschrijver et al., 2016; Nijhof et al., 2016). The timing properties differ from the version used by Kovàcs et al. (2010; see further). The storyline remained the same.

The participants watched short video animations of 720 × 480 pixels. Each video consisted of a *Belief Formation phase* and an *Outcome phase*. During the movies, the beliefs of the agent (A) and the participant (P) about the presence of a ball were manipulated ('+' if the ball was present behind the occluder, '−' if the ball was absent). The movie always begins with Buzz Lightyear entering the scene and placing a ball on the table in front of an occluder. The ball then starts moving and rolls behind the occluder. From that moment on, four different scenarios, depending on the experimental condition, were possible:

1. *True belief* Positive content condition (Participant+Agent+)

   The ball rolls from behind the occluder halfway to the right and then back behind the occluder. The agent walks out of the scene with the (implicit) belief that the ball is behind the occluder (A+). In the absence of Buzz, the ball starts rolling halfway to the right and then halts behind the occluder, resulting in the participant holding the same belief as Buzz (P+).
2. *False belief* Negative content condition (P+A−)

   The ball rolls from behind the occluder halfway to the right and then out of the scene. The agent walks out of the scene with the (implicit) belief that the ball is not behind the occluder (A−). In the absence of Buzz, the ball rolls back into the scene and halts behind the occluder. The result is that the participant now holds a different belief than Buzz (P+).
3. *False belief* Positive content condition (P−A+)

   The ball rolls from behind the occluder halfway to the right and then back behind the occluder. The agent walks out of the scene with the (implicit) belief that the ball is behind the occluder (A+). In the absence of Buzz, the ball starts rolling halfway to the right and then out of the scene. The result is that the participant now holds a different belief than Buzz (P−).
4. *True belief* Negative content condition (P−A−)

   The ball rolls from behind the occluder halfway to the right and then out of the scene. The agent walks out of the scene with the (implicit) belief that the ball is not behind the occluder (A−). In the absence of Buzz, the ball rolls back into the scene and then again out of the scene, resulting in the participant holding the same belief as Buzz (P−).

In the *Outcome phase*, Buzz walks back into the scene and the occluder falls. This reveals whether the ball is or is not present behind the occluder. The ball was present in half of the trials. The presence of the ball was completely random and independent of the belief formation phase. Participants were asked to press a key ('V' on the keyboard) when the agent left the scene and to press a key when the occluder fell and a ball was present ('B' on the keyboard). As a result of combining the *Belief Formation* and *Outcome* phase, there were eight different movies/conditions that were each shown ten times. Therefore, the entire experiment consisted out of 80 trials, presented in a randomized order in two blocks of 40 trials with a short break in between. A blank screen was shown for 2000 ms before every new trial (intertrial interval; ITI). Prior to the start of the experiment, four practice trials were presented to the participants. They received feedback on their response after each of the respective practice trials.

Whereas in the original task (Kovacs et al., 2010) the timing differs across conditions (specifically, the moment when the agent left the scene), in our adjusted version, equal timing of events during the task was assured (see Fig. 1).

Every trial consists out of different movie clips and frames that are presented after each other as one movie that lasts 12,868 ms using Presentation Software, version 18.1 (Neurobehavioral Systems Inc., San Francisco, CA).

Questionnaires

Two questionnaires were administered to measure ASD symptomatology: the Autism Spectrum Quotient (AQ; (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001; Hoekstra, Bartels, Cath, & Boomsma, 2008) and the Social Responsiveness Scale-Adult Version (SRS-A; Constantino & Gruber, 2005).

The Autism Spectrum Quotient (AQ) The AQ is a self-report screening questionnaire with 50 items assessing autistic traits in adults. It results in a score on five subscales with ten items each: social skills, attention switching, attention to detail, communication and imagination. Each item can be scored on a scale from 1 to 4, and later, this score was converted to a dichotomous outcome (0 or 1).

The Social Responsiveness Scale-Adult Version (SRS-A) The SRS-A is a self-report questionnaire for adults between 18 and 65 years old that measures behavioral dimensions that are characteristic to ASD. It consists of 64 items that can be scored on a scale from 1 to 4 and results in four subscales: social awareness, social communication, social motivation and rigidity/repetitiveness.

Procedure

The study consisted out of two experimental tasks (here, we focus on the Buzz Lightyear task, and the other task will be discussed elsewhere) and three pen-and-paper questionnaires (two of which are described above). After participants signed the informed consent, they were asked to fill out the SRS-A (Constantino & Gruber, 2005). Subsequently, one

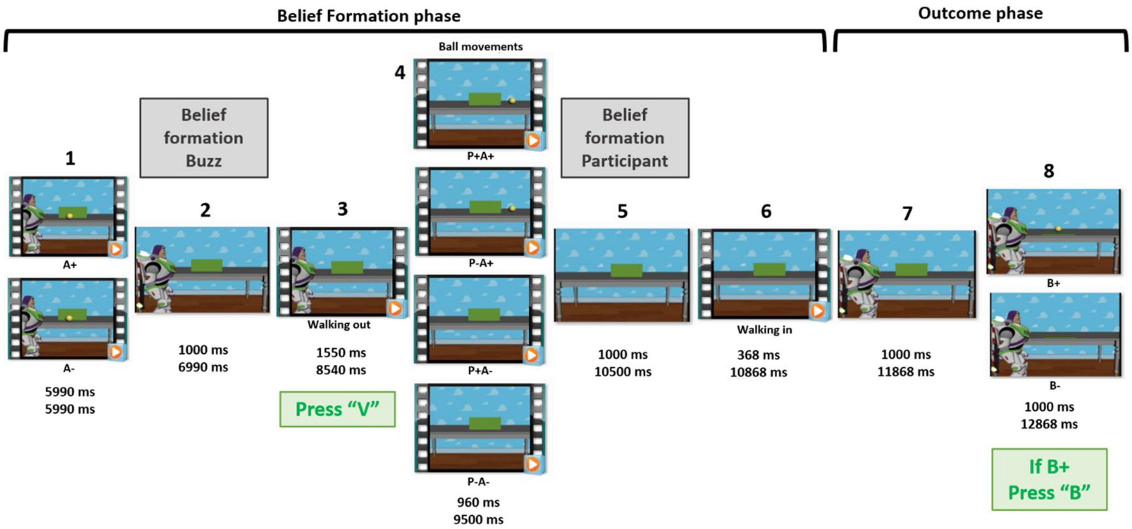

Fig. 1 Design of the paradigm: The timing is consistent across all conditions. The durations of every movie clip and frame are shown in the first row, and the second row shows the cumulative durations throughout the trial. Each frame is presented 1000 ms to make it possible to time lock certain events (i.e., belief formation Buzz, belief formation participant, possible conflict between the two beliefs, outcome). The instructions that were given to the participants are indicated in green. Total duration of one trial is 12,868 ms, and there was a fixed intertrial interval of 2000 ms

of the experimental tasks was presented and after completion they filled out a Dutch debriefing form with five questions based on the one used by Schneider et al. (2013). This was followed by the AQ (Baron-Cohen et al., 2001), another experimental task was introduced, also followed by a debriefing form, and finally, the last questionnaire, namely the Toronto Alexithymia Scale (TAS-20; Bagby, Parker, & Taylor, 1994), was administered. The order of the experimental tasks was counterbalanced, and there were no order effects. The entire session took approximately 1 h. At the end of the session, participants were awarded a course credit and to encourage motivation a monetary bonus was given to the top five students with the best accuracies and reaction times on both experimental tasks.

Statistical analysis

Reaction times for detection of the ball were recorded at the end of each trial in which the ball was present behind the occluder. Outlier analysis was carried out. All responses more than three standard deviations above or below the participants overall mean or less than 100 ms were removed from analyses. This resulted in a loss of 46 data points (2.05%) across all 57 participants. Control analyses revealed that removing the outliers did significantly change any of the findings.

To investigate the effect of own belief and agent belief on reaction times, a 2-by-2 repeated measures ANOVA was carried out with own belief (P+, P−) and agent belief (A+, A−) as within-subject factors. We controlled for multiple comparisons by means of a Bonferroni correction. In case of violation of the assumption of sphericity, Greenhouse–Geisser or Huynh–Feldt correction was applied. Additionally, we performed a planned comparison for the difference between P−A− and P−A+, referred to as the ToM-index, which was taken as evidence for spontaneous ToM in the original paper of Kovàcs et al. (2010) and was later criticized by Philips et al. (2015). Finally, Pearson correlations were performed to explore the link between the ToM-index and ASD symptomatology as measured with the AQ and the SRS-A. If there were missing data for the SRS-A, the median of that scale was used as the score on that item (Constantino & Gruber, 2005). Missing items on the AQ were estimated using the expectation–maximization technique. In the correlational analyses, we used scaled scores for the ToM-index ((RT P−A−) − (RT P−A+)/(RT P−A−) + (RT P−A+)), to control for potential confounding effects of differences in overall RT. All statistical analyses were carried out with IBM SPSS statistics (SPSS Inc., Chicago, IL, USA).

Results

Behavioral results

Accuracy Participants were asked to press a key ('B') when the ball was present behind the occluder in the outcome phase. Only a few omission errors were made (1.27% of trials), and hence, these were not further analyzed. In 8.37% of the trials, they pressed the key when the ball was not present behind the occluder (i.e., 'false alarms'). There is no effect of condition on the number of false alarms (F (3,168) = 1.63, $p = 0.19$, $\eta_p^2 = 0.03$; sphericity assumed).

Reaction time The mean reaction time to the ball was 351 ms (SD = 8.85 ms; 95% CI [333.34, 368.81]). Figure 2 displays the mean reaction time for each condition.

To investigate the effect of own belief and agent belief on reaction times, a 2 × 2 repeated measures ANOVA was performed. A significant main effect of own belief (F(1,56) = 59.09, $p < 0.001$, $\eta_p^2 = 0.51$) and agent belief (F(1,56) = 6.50, $p = 0.01$, $\eta_p^2 = 0.10$) was found. There was no significant interaction effect of own belief and agent belief (F(1,56) = 0.62, $p = 0.44$, $\eta_p^2 = 0.01$). Taking a closer look at the main effect of own belief, we observed that participants respond 29.59 ms faster (SD = 3.85, 95% CI [− 37.29, − 21.87], $p < 0.001$) to the presence of the ball in the conditions were the participants hold the belief that the ball would be present (P+ conditions) compared to the conditions were the participants belief the ball would be absent (P− conditions). The main effect of agent belief indicates that the belief of the agent about the presence or absence of the ball also influences how fast the participant detects the ball. Participants are 9.14 ms faster (SD = 3.58, 95% CI [− 16.32, − 1.96], $p = 0.01$) in the conditions were the agent holds the belief that the ball would be present (A+ conditions) compared to the conditions were the agent had the opposite belief (A− conditions).

The planned comparison between P−A− and P−A+ (the ToM-index) shows that participants were significantly slower in the P−A− condition compared to the P−A+ condition ($p = 0.03$; 95% CI [1.38, 22.80]). As this was a planned comparison based on the findings of others (Deschrijver et al., 2016; Kovàcs et al., 2010; Nijhof et al., 2016), no Bonferroni correction was applied. Table 1 shows all comparison between conditions, with Bonferroni correction for the other comparisons (p value of 0.0083 (0.05/6)). All conditions differ significantly, even after Bonferroni correction, from each other with the exception of P+A+ and P+A−.

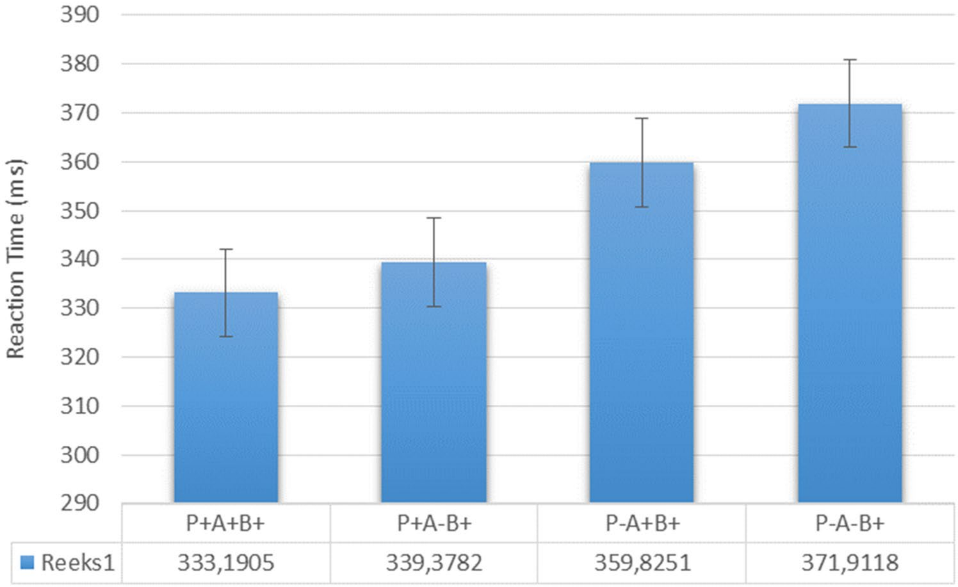

Fig. 2 Mean reaction times (ms) per condition

Table 1 Overview of the statistical comparisons between the four conditions

Condition 1	Condition 2	Mean diff. (1–2)	Std. error	95% CI	p value
P−A−	P−A+	12.09	5.35	1.38–22.80	0.028*
	P+A−	32.53	6.27	19.97–45.09	< 0.001*
	P+A+	38.72	4.75	29.21–48.23	< 0.001*
P−A+	P+A−	20.45	5.72	8.98–31.91	0.001*
	P+A+	26.64	4.31	18.00–35.27	< 0.001*
P+A−	P+A+	6.19	5.04	− 3.90–16.28	0.224

P = Participant, A = Agent, **+** = belief ball is present, − = belief ball is absent. The colored row indicates the planned comparison (P−A− vs. P−A+)

*Indicates a significant effect

Questionnaires

All 57 participants filled out both questionnaires. For the SRS-A, the scores ranged between 16 and 109 with a mean score of 41.39 (SD = 16.23). There was 0.003% missing data (10 of 3420 items). The scores on the AQ varied between 8 and 31 with a mean score of 15.14 (SD = 5.43). There was 0.003% missing data (9 of 2850 items). As expected, these two questionnaires are strongly correlated ($r = 0.67$, $p < 0.001$). However, neither the SRS-A nor the AQ correlated significantly with the ToM-index ($r = -0.08$, $p = 0.54$ for the SRS-A and $r = -0.03$, $p = 0.84$ for the AQ).

Discussion

Kovács et al. (2010) developed a ball detection paradigm to measure spontaneous mentalizing. They observed that when participants do not expect the ball to be present (P−), they detect the ball faster when the agent believes the ball is present (A+) rather than absent (A−), which they took as evidence for spontaneous mentalizing (Kovács et al., 2010). After that, several studies applied this paradigm and were able to replicate this finding (Bardi et al., 2017; Deschrijver et al., 2016; Nijhof et al., 2017). However, Phillips et al. (2015) argued that the observed effects are possibly not the result of mentalizing abilities but rather are the outcome of timing artifacts within the paradigm, more precisely differences in timing between conditions for the attention check, referred to as the attention-check hypothesis. The objective of the current study was to investigate if the ToM-effect would still be present when all timings would be equalized, thus eliminating any potential timing artifact. The ToM-index was observed, and hence, this effect cannot be explained by a timing confound.

We found that the participants reactions were influenced by both their own belief and the agent belief as reflected in significant main effects of own belief and agent belief on reaction times. This shows that the participant's reaction times are not only affected by their own belief, but also by the belief that is attributed to an agent. As the original work of Kovács et al. (2010), we focused on the difference between P−A+ and P−A− conditions, which is referred to as the ToM-index, and was later contested by Phillips et al. (2015). We performed a planned comparison for this effect. In line with previous work, we found a significant ToM-index: Participants were significantly faster in the condition where the agent believes the ball will be present (P−A+), compared to the condition where the agent believes the ball is absent while the participant believes the ball would be absent (P−A−). This implies that even if they believe the ball will be absent, they spontaneously take the belief of the agent into account, enabling them to respond faster to the presence of the ball. These findings indicate that the attention-check hypothesis of Phillips et al. (2015) do not explain the results that are found with this paradigm. They argued that the effects were due to a shorter delay between the attention-check key press and the ball detection response in the P+A+ and P−A− conditions (Phillips et al., 2015). Differences in reaction time are then the result of the psychological refractory period. However, the current study refutes this as all possible timing artifacts are removed from the paradigm; thus, the found effects cannot be explained by the attention-check hypothesis.

Interestingly, we did not only observe a significant ToM-index, but also a main effect of agent belief was found. This effect was hypothesized, but not observed in the original study of Kovács et al. (2010). Indicating that, our findings provide even stronger support for spontaneous mentalizing, compared to only the ToM-index. Most studies focused specifically on the ToM-index; however, a main effect of the other agent belief has been reported in another study using the paradigm with 'confounded' timing (Nijhof et al., 2017) as well, and hence, we can only speculate about the reasons for our finding. It may be that the effect of the agent belief on ball detection is (somewhat) stronger when oneself does not have a representation of the ball (P−A− vs P−A+) than when there already exists a representation of the ball (P+A− vs P+A+), which may explain the inconsistency across studies. Future well-powered studies are needed to further address this issue.

Phillips et al. (2015) highlighted that we need to get a better understanding of (spontaneous) mentalizing processes, but that this requires valid paradigms to test these aspects of social cognition. The current study demonstrates that the effects found with the paradigm of Kovács et al. (2010) are not solely due to timing confounds. Several studies (e.g., Bardi et al., 2017; Deschrijver et al., 2016; Nijhof et al., 2016, 2017) successfully employed this paradigm and contested the attention-check hypothesis with theoretical and statistical arguments. Nevertheless, to completely rule out the attention-check hypothesis, the timing difference between conditions needs to be eliminated. Phillips et al. (2015) made an attempt by putting the attention check at the moment when the agent comes back in the scene (same in every condition) and by removing the attention check. Nevertheless, these approaches ignore the true purpose of the attention check, namely making sure that participants pay attention to the agent at the crucial moment when he forms a belief about the location of the ball, without the participant being aware of this purpose.

As of yet, no study had assured fully equal timing between events in this paradigm (while also taking the purpose of the attention check into account), which is needed to fully rule out any potential timing confounds. The current study provides empirical evidence because here we show that the

ToM-index cannot be explained in terms of differences in timing of events or refractory period. This conclusion is in accordance with recent fMRI findings indicating that core ToM regions, such as the TPJ and the medial prefrontal cortex (mPFC), are activated when using the spontaneous ToM task (Bardi et al., 2017; Nijhof et al., 2018). This strengthens the conclusion that findings cannot solely be attributed to confounds but do reflect spontaneous mentalizing processes.

To assure that belief attribution was spontaneous and not explicit, participants were asked to fill out a debriefing form (see also, Schneider et al., 2013). This debriefing procedure revealed that participants were not consciously tracking the belief of the agent. Participants were also asked if they were paying attention to the movements of the ball. This is important to take into account because if the participants do not pay attention to the movements of the ball, and thus are not tracking its location, they cannot attribute a belief about its location to another person. Nine of the participants reported that they only focused on the outcome (the falling of the green occluder) and that they did not track the movements of the ball. Excluding these nine participants did not change the effects. Exploratory analyses on this very small sample suggest that when there is no attention to the ball, the reaction times only differ between the P+ and P− conditions, and that no agent effect nor the ToM-index is observed. Our sample is too small to draw any strong conclusions, but this notion may be of importance for future studies. Two things can be considered. First, it indicates the importance of the attention check. Second, future research should be aware of this issue and include a debriefing to control if participants actually paid attention to the object (here a ball). This may be important to assure that results and group differences in future studies may not be due to differences in paying attention to the ball.

As an additional aim of our study, we investigated the relation between ASD symptomatology as a dimensional trait in a neurotypical population and spontaneous ToM, but no correlation was found. This finding is in line with a recent study by Nijhof et al. (2016) and can be due to restricted variance in ASD symptomatology in our neurotypical sample. In a study with neurotypical adults with higher or lower levels on ASD symptomatology (a priori selected based on scores on the short AQ), a group difference was found: Neurotypicals with lower levels of ASD symptomatology show a significant ToM-index, while this was absent in the group with higher levels of ASD symptoms (Nijhof et al., 2017). Deschrijver et al. (2015) also found a correlation between ASD symptomatology and spontaneous ToM (the ToM-index), however, only in their ASD sample. Hence, overall findings do suggest a negative association between ASD symptoms and spontaneous mentalizing, especially when ASD symptoms are more severe but more research is definitely warranted, as findings are not conclusive.

In summary, Phillips et al. (2015) argued that the found effects with the paradigm developed by Kovács et al. (2010) were due to timing artifacts caused by the psychological refractory period and that therefore the paradigm is not suited to measure spontaneous mentalizing. Here, we showed that when keeping the purpose of the attention check intact and equalizing the timings of all events across conditions, we found significant effects of the agent belief on ball detection speed and could replicate the initial finding of Kovács et al. (2010); we observed a significant ToM-index. Our results refute the attention-check hypothesis and suggest that we spontaneously track other agents' beliefs within this paradigm. No association with ASD symptoms in our neurotypical sample was found, warranting further research.

Acknowledgements This work was supported by the Special Research Fund of Ghent University (project number BOF.DOC.2015.0096.01).

Funding This work was funded by the Special Research Fund of Ghent University (project number BOF.DOC. 2015.0096.01).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in this study involving human participants were in accordance with the ethical standards of 'The Ethical Committee of the Faculty of Psychology and Educational Sciences of Ghent University' and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

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

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research, 38,* 23–32.

Bardi, L., Desmet, C., Nijhof, A., Wiersema, J. R., & Brass, M. (2017). Brain activation for spontaneous and explicit mentalizing in adults with autism spectr. *Social Cognitive and Affective Neuroscience, 12*(3), 391–400. https://doi.org/10.1093/scan/nsw143.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism Spectrum Quotient: Evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17. https://doi.org/10.1023/A:1005653411471.

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

Constantino, J. N., & Gruber, C. P. (2005). *Social responsiveness scale (SRS)*. Los Angeles, CA: Western Psychological Services.

Deschrijver, E., Bardi, L., Wiersema, J. R., & Brass, M. (2016). Behavioral measures of implicit theory of mind in adults with high functioning autism. *Cognitive Neuroscience, 7*(1–4). https://doi.org/10.1080/17588928.2015.1085375.

Frith, U. (2012). Why we need cognitive explanations of autism. *Quarterly Journal of Experimental Psychology, 65*(11), 2073–2092. https://doi.org/10.1080/17470218.2012.697178.

Hill, E. L., & Frith, U. (2003). Understanding autism: Insights from mind and brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1430), 281–289. https://doi.org/10.1098/rstb.2002.1209.

Hoekstra, R. A., Bartels, M., Cath, D. C., & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the autism-spectrum quotient (AQ): A study in Dutch population and patient groups. *Journal of Autism and Developmental Disorders, 38*(8), 1555–1566. https://doi.org/10.1007/s10803-008-0538-x.

Kovács, Á. M., Kühn, S., Gergely, G., Csibra, G., & Brass, M. (2014). Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PLoS One*. https://doi.org/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*(2010), 1830–1834. https://doi.org/10.1126/science.1190792.

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

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

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

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

Nijhof, A. D., Brass, M., & Wiersema, J. R. (2017). Spontaneous mentalizing in neurotypicals scoring high versus low on symptomatology of autism spectrum disorder. *Psychiatry Research, 258,* 15–20. https://doi.org/10.1016/j.psychres.2017.09.060.

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

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences, 4,* 515–526.

Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism. *Brain and Cognition, 55*(1), 209–219. https://doi.org/10.1016/j.bandc.2003.04.002.

Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition, 129*(2), 410–417. https://doi.org/10.1016/j.cognition.2013.08.004.

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

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

Schuwerk, T., Vuori, M., & Sodian, B. (2015). Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. *Autism, 19*(4), 459–468. https://doi.org/10.1177/1362361314526004.

Senju, A. (2013a). Atypical development of spontaneous social cognition in autism spectrum disorders. *Brain and Development, 35*(2), 96–101. https://doi.org/10.1016/j.braindev.2012.08.002.

Senju, A. (2013b). Spontaneous theory of mind and its absence in autism spectrum disorders. *Brain and Development, 35*(2), 96–101. https://doi.org/10.1177/1073858410397208.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.