This is a repository copy of *The effect of social presence on mentalizing behavior*.

Article:

The Effect of Social Presence on Mentalizing Behavior

Emma J. Morgan, Daniel J. Carroll, Constance K. C. Chow, Megan Freeth

Department of Psychology, The University of Sheffield

Abstract

Our behavior is frequently influenced by those around us. However, the majority of social cognition research is conducted using socially isolated paradigms, without the presence of real people (i.e., without a "social presence"). The current study aimed to test the influence of social presence upon a measure of mentalizing behavior in adults. Study 1 used a first-order theory of mind task; and study 2 used a second-order theory of mind task. Both studies included two conditions: *live*, where the task protagonists were physically present acting out the task, or *recorded*, where the same task protagonists demonstrated the task in a video recording. In both experiments, participants were affected by the social presence and demonstrated significantly different patterns of behavior in response to the presence of real people. This study, therefore, highlights the critical importance of understanding the effect of a social presence in mentalizing research, and suggests that the inclusion of a social presence needs to be given strong consideration across social cognition paradigms.

Keywords: Social cognition; Mentalizing; Social presence; Theory of mind

1. Introduction

As humans, we are a highly social species and, as such, are predisposed to attend to social information within our environment. This preference emerges from an early age (Simion, Regolin, & Bulf, 2008) and continues into adulthood (Birmingham & Kingstone, 2009). Of key interest, much research has demonstrated that participants are sensitive to the social nature of stimuli used within lab experiments (Gobel, Kim, & Richardson, 2015; Risko, Laidlaw,

Correspondence should be sent to Emma J. Morgan, Department of Psychology, The University of Sheffield, Sheffield S1 2LT, UK. E-mail: emma.morgan@sheffield.ac.uk

Freeth, Foulsham, & Kingstone, 2012; Risko, Richardson, & Kingstone, 2016). Therefore, a current concern within social cognition research relates to the appropriateness of social cognition paradigms conducted with socially isolated stimuli—that is, stimuli presented with no real person present—such as cartoons or images. Even when stimuli are rich and dynamic (such as videos of real people), they have still been found to elicit different patterns of behavior than when completing the same task in the presence of real people (hereafter termed as "social presence") (Reader & Holmes, 2016). Such behavioral changes can range broadly from differences in participants' gaze behavior (social attention) to differences in overall task performance, differences in attempts at reciprocal social interaction, or differences in mimicking actions. This difference is found even if the stimuli presented to the participants are exactly identical. For example, when watching the same Skype conversation, people demonstrate different patterns of interactive behavior when they believe the stimulus to depict a person in real time than when they believe the same stimulus to be a recording (Cole, Skarratt, & Kuhn, 2016). This, therefore, suggests that experiments featuring a social presence can lead to different results to experiments without a social presence, and that these differences should be a critical consideration in social cognition research. However, to date, this research has been limited to a few specific areas of social cognition (such as social attention). No systematic investigations have yet been conducted to investigate whether social presence impacts other core areas of social cognition, such as mentalizing behavior.

Within social cognition research, in general, there is an increased focus on understanding how cognitive mechanisms differ when measured in response to a social presence, rather than in isolation (Cole et al., 2016; De Jaegher, Di Paolo, & Gallagher, 2010; Reader & Holmes, 2016). A key proposal is that social environments are not merely contexts in which behavior happens; rather, they are a critical component of human cognitive processing, and can drive and instigate cognitive mechanisms in their own right (De Jaegher et al., 2010). Interestingly, behavioral changes have been shown to occur even when a participant is made aware that a social partner will not engage in a social interaction; it appears that the mere potential for an interaction is enough to drive these changes in behavior (Gregory et al., 2015). Further, this phenomenon has been detected at the cortical level, whereby simply viewing a social interaction between others leads to increased cortical excitability—even if there is no potential for the participant themselves to interact (Aihara, Yamamoto, Mori, Kushiro, & Uehara, 2015). This suggests that the cognitive processes that are engaged in response to a social presence are not just limited to our own personal interactions, but also encompass another person's potential interactions with another social partner (Reader & Holmes, 2016).

The ability to recognize that others may hold desires, beliefs, and intentions that differ to our own is termed "theory of mind" or "mentalizing" (Premack & Woodruff, 1978). This skill is critical within our everyday lives yet, surprisingly for a construct rooted in social behavior, to date, there has been little consideration regarding the effect of social presence in mentalizing paradigms. A limited number of studies provide key evidence that this is likely to be an important factor to consider. For example, Chevallier et al. (2014) found that when children completed false-belief tasks, their performance tended to be significantly better when instructions were given verbally by an experimenter, than when instructions were read by participants themselves. Redcay and Schilbach (2019) discuss evidence that, at the neural level,

parts of the mentalizing network engage in response to a real social partner even when completing tasks with no mentalizing demands; but that these areas do not show similar levels of engagement for identical scenarios with recorded stimuli. Additionally, preliminary research indicates that the inclusion of naturalistic social scenarios promotes mentalizing behavior in nonhuman primates (Krupenye, Kano, Hirata, Call, & Tomasello, 2016). However, this effect is yet to be fully investigated in humans, with only one study to date demonstrating that naturalistic stimuli can lead to a slight improvement in mentalizing performance (although this improvement was not statistically significant: Kulke, Wubker, & Rakoczy, 2019). While research has, therefore, begun to yield new and promising discoveries regarding the importance of social presence in mentalizing research, there is still little understanding regarding the behavioral impact of a social presence in the tasks used to measure mentalizing behavior.

The aim of the present study, therefore, was to determine the extent and nature of the effect of a social presence upon measures of mentalizing behavior in adults. This study seeks to investigate the impact of a social presence with regard to two different forms of mentalizing behavior: "implicit" and "explicit" mentalizing. The terms "implicit" and "explicit" mentalizing can vary in how they are operationalized; for clarity, this paper will use the terms as previously defined by Gawronski and Hahn (2019), and Rosenblau, Kliemann, Heekeren, and Dziobek (2015). An "implicit" mental representation is characterized by a participant's response being automatic, unintentional, and independent of cognitive resources. In contrast, an "explicit" mental representation is characterized by a participant's response being intentional and dependent on cognitive resources. Therefore, we will apply the terms "implicit" and "explicit" to the cognitive processes we are seeking to measure. These processes will be measured using two different behavioral measures. To measure implicit mentalizing behavior, we will use an eye-tracking paradigm, in which the participant is unaware of what behavior is being measured while completing the task. To measure explicit mentalizing behavior, we will use a behavioral paradigm in which we ask participants to provide a direct, conscious response (a key press or finger point) in response to the task.

The suggestion that social presence could elicit different patterns of behavior in mentalizing tasks is a highly plausible consideration. Mentalizing has typically been studied using "false belief" tasks. Such tasks assess a participant's ability to understand that a protagonist holds a false belief about the location or status of an object (Wimmer & Perner, 1983). However, the method of administration of false belief tasks can differ considerably across studies. While some studies use static, schematic images (e.g., Baron-Cohen, Leslie, & Frith, 1986; Van der Wel, Sebanz, & Knoblich, 2014), others use dynamic, naturalistic stimuli (Schulze & Tomasello, 2015). These paradigms clearly differ in the complexity of the stimuli used, and yet the conclusions drawn from each experiment are broadly accepted as analogous (Cole et al., 2016). This study will, therefore, investigate the effect that social presence has upon a measure of mentalizing behavior in adults, by comparing performance on a mentalizing task conducted with either prerecorded stimuli or with live task protagonists.

2. Study 1

The overall aim of this research was to investigate whether social presence, defined as the presence or absence of task protagonists, influences participants' behavior on mentalizing tasks. The purpose of study 1 was to pilot and test the proposed research design, and to inform the sample size selection for study 2. In study 1, participants completed four versions of a first-order theory of mind task, in which a protagonist had to keep track of an object's location. The task provided measures of behavior associated with both implicit *and* explicit mentalizing. These processes were measured using two variations of the same task, featuring two different behavioral measures (Eye-Tracking or Direct Response). Tasks were completed in one of two conditions: in the "Live" condition, the task protagonists were physically present acting out the task trials, and in the "Recorded" condition, the same task protagonists were shown via a video recording. In all other respects, the two conditions were identical. A first-order theory of mind task was considered appropriate, as this type of false-belief task has been commonly used within mentalizing research and has previously elicited eye-movement behavior associated with implicit mentalizing in adults (Schneider, Bayliss, Becker, & Dux, 2011).

With the use of a first-order of theory of mind task, it was expected that participants would show a high level of accuracy in each condition. However, no previous studies have compared a measure of implicit mentalizing between Live and Recorded conditions; hence, we were interested to investigate whether eye-movement behavior recorded via eye-tracking technology would differ in the presence of live task protagonists compared to when no live protagonists were present. If there were differences in eye-movement behavior between the Live and Recorded conditions, this would suggest that the context in which we attempt to measure mentalizing has important implications for the reliability of the collected data. This would indicate that social presence is a crucial factor that must be taken into account when creating paradigms designed to investigate mentalizing.

3. Method

3.1. Participants

Thirty-one participants (19 females and 12 males), with a mean age of 22 years (range = 19–25, *SD* = 3.12) took part in the study. Undergraduate Psychology students received course credit for taking part. The study was approved by the Department of Psychology Ethics Committee, and all participants gave informed consent before participating. All participants had normal, or corrected to normal, vision.

3.2. Design and data handling

The study used a within-subjects design with two independent variables: condition (Live or Recorded) and task type (Eye-Tracking or Direct Response). Each participant, therefore, completed four versions of the tasks: (1) Live-Eye-Tracking; (2) Live-Direct Response; (3)

Recorded-Eye-Tracking; and (4) Recorded-Direct Response. The order in which the Live and Recorded conditions were presented was counterbalanced between participants, but the Eye-Tracking task was always completed before the associated Direct Response task. The study paradigm was a nonverbal first-order theory of mind task, which required participants to track the knowledge state of two protagonists. In the Live conditions, a sequence was acted out in real time in front of the participant; in the Recorded conditions, the participant was shown a video of the same sequence presented on a computer screen.

In each version of the task, the participant watched as two protagonists (the "hider" and the "seeker") acted out a role-play in which the hider hid a ball, and the seeker tried to find the ball (Fig. 1). On each trial, the seeker would leave the scene, and the hider would then manipulate two visually identical boxes—either swapping their positions, or picking them up and replacing them in the same positions. On a number of trials, the seeker would peek back and watch the scene, meaning that on those trials, they could accurately track the ball's location. On other trials, the seeker did not watch the hider's actions, and thus could not reliably track the location of the ball. Participants were told in advance that the hider would not know whether the seeker peeked on each trial, and that the seeker would not know whether the hider would swap the boxes.

To ensure that the conditions were as closely matched as possible, the videos used in the Recorded condition were created from recordings of the Live condition; the actors in the recording were the same people who enacted the live display. The sole difference between each condition was the live or recorded nature of the scene. During the Live version of the scene, the actors avoided eye contact, physical contact, or any actions that could be construed as gestures to the other confederates while acting out the task. They also avoided any actions that could be construed as gestures to the participant. Actors were instructed to maintain a neutral facial expression throughout, and were instructed to avoid making any eye contact with the participant, and instead to direct their gaze at a central point located away from the participant's location. Finally, the seeker was instructed not to look toward any of the boxes when returning to the scene, and instead to maintain their gaze toward a central position directly in front of them.

For the Eye-Tracking task, participants were instructed simply to watch the role-plays; during these tasks, participants' eye movements were recorded. The Eye-Tracking trials measured whether participants spontaneously tracked the seeker's false belief. This was measured both via the participants' anticipatory gaze behavior and by their overall pattern of preferential looking. Attending to the box congruent with the seeker's false belief (i.e., the box where the seeker would search, based on their knowledge of what they had seen) was considered to be the "correct" response. The anticipatory gaze data indicated which box the participant first fixated; if the participant's first fixation was to the correct box, then the fixation was labeled as a correct response. To allow comparison with previous literature, a second Eye-Tracking measure was also calculated. In order to assess the participants' preferential looking patterns, a preferential looking score was calculated for each trial. This was computed by taking the total duration of fixations on the correct box, and then dividing this by the total duration of fixations on all boxes. Trials in which the participant did not direct any fixations toward the four boxes were removed prior to analysis. The preferential looking score for each trial was
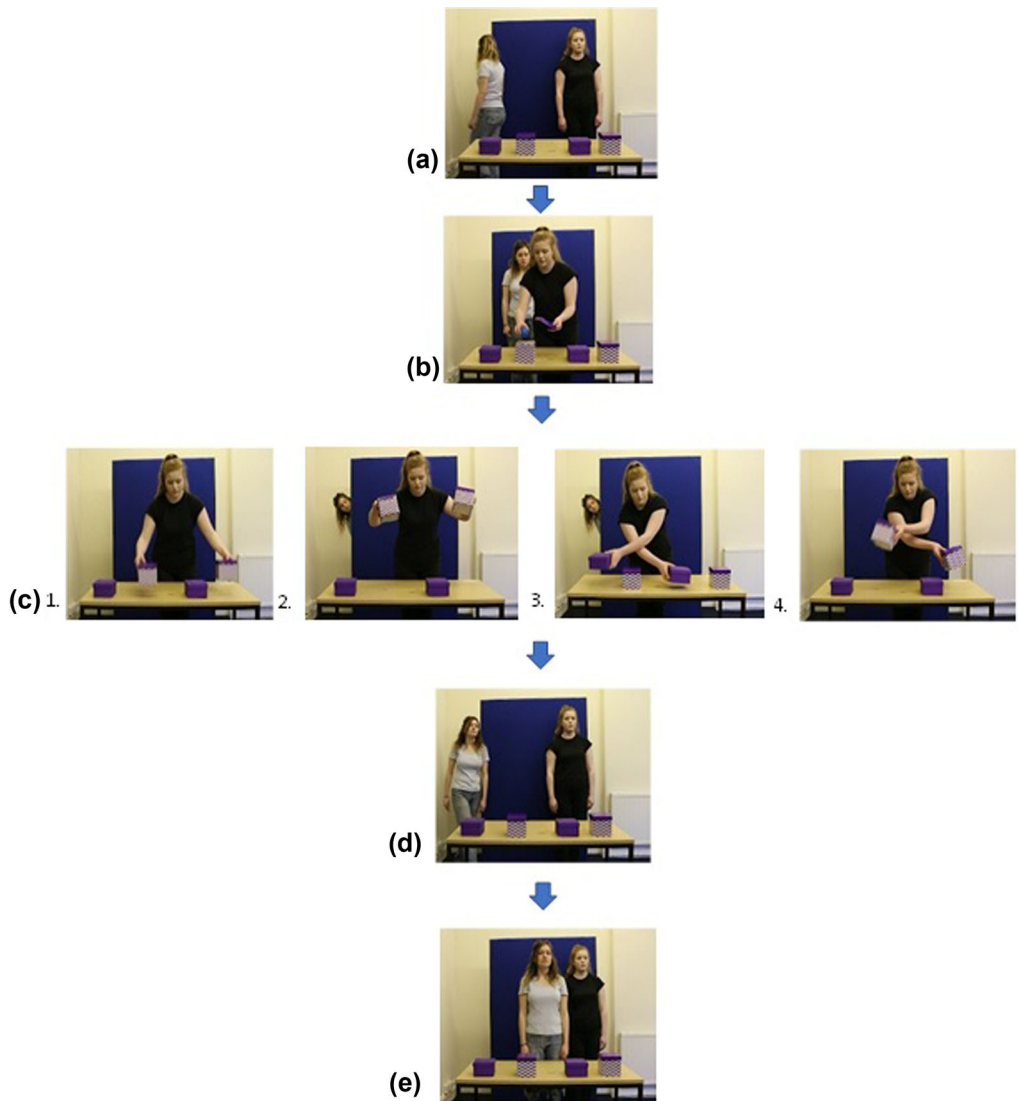
Fig. 1. The experimental procedure. The trial type is labeled based on the seeker's belief. (a) The hider hides the ball, watched by the seeker. (b) The seeker leaves. (c). (1) True-belief trial: the hider lifts the boxes and replaces them in their original positions; the seeker remains hidden. (2) True-belief trial: the hider lifts the boxes and replaces them in their original positions; the seeker watches. (3) True-belief trial: the hider swaps the boxes, the seeker watches. (4) First-order false-belief trial: the hider swaps the boxes, the seeker remains hidden. (d) The seeker emerges to search for the ball. (e) The seeker remains standing behind the table. During the explicit trials, the participant provides a response.

then averaged across the total number of trials to give an overall score for each participant. This measure, therefore, gave an index of whether participants preferentially looked toward the correct location over the course of the experiment. In order to determine if participants demonstrated a systematic bias to attend to the correct box, both the preferential looking data and the first fixation data were analyzed to test whether performance on at least one measure was significantly greater than would be expected from a random allocation of attention. The threshold used to identify systematic attendance to the correct box was determined to be 25%, based on the potential for participants to randomly allocate their gaze to any of the four boxes used in the paradigm. To establish whether participants preferentially attended more to the correct box than any of the other three boxes, we calculated the proportion of time participants spent looking at each of the four boxes on each trial, and then compared the value of the correct box against the threshold of 0.25 (25%). For the first fixation data, performance was based on which box, out of the four boxes, was first fixated.

To check whether participants were engaging in implicit mentalizing processes to complete the task, each participant completed a funneled debriefing procedure at the end of the testing period (see Appendix A). This procedure was adapted from that used by Schneider, Slaughter, Bayliss, and Dux (2013), which was in turn adapted from a procedure designed to test higher implicit processes by Bargh and Chartrand (2000). The debriefing procedure consisted of a series of questions that probed the participants' understanding of the task, and thus could confirm whether any participants had engaged in conscious belief processing for the Eye-Tracking version of the task.

The Direct Response task followed the same procedure as the Eye-Tracking task. However, at the end of each trial, participants were asked to indicate which box they thought the seeker would look in. Responses were recorded via a finger point (for the Live condition) or by a button press (for the Recorded condition). In the Live condition, the finger point was captured via the cameras on a mobile eye tracker; these were replayed by the experimenter and manually coded for accuracy. For the Eye-Tracking trials in the Recorded condition, the button press was recorded as either correct or incorrect for each trial during the experiment.

3.3. Apparatus

Participants' eye movements were recorded during each experimental condition. SMI (Senso Motoric Instruments) mobile eye-tracking glasses were used during the Live condition, and a desktop-mounted EyeLink1000 eye-tracker was used during the Recorded condition. The mobile eye-tracking glasses were calibrated using a one-point calibration, which required participants to focus on a point within their visual field (in this case, the tip of the experimenter's extended finger), at a distance within the range that the main trial action took place. Accuracy of the calibration was monitored in real time during the experiment, through observation of the participants' gaze location in context of the visual field recording. As each trial unfolded, the cursor representing the participants' gaze was checked to ensure that it matched up with points of interest (such as the blue ball or the red clue; see Fig. 6). The glasses had an eye-tracking range of 80° horizontal, 60° vertical, with a binocular 30 Hz temporal and up to 0.1° spatial resolution combined with 24 Hz front-view camera with a field
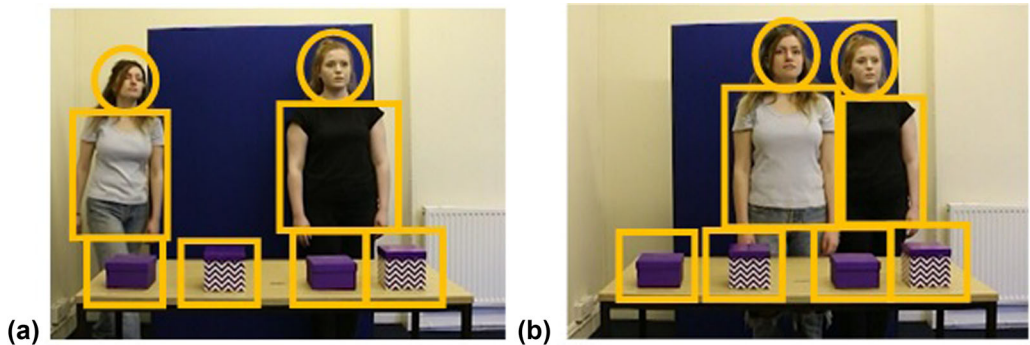
Fig. 2. The areas of interest for the anticipatory gaze data analysis. (a) At the start of the interest period, as the seeker emerges from behind the screen. (b) At the end of interest period, while the seeker remains behind the table.

of view 60° horizontal and 46° vertical. The desktop EyeLink1000 had an eye-tracking range of 32° horizontal, 25° vertical, with a monocular 1000 Hz temporal and up to 0.01° spatial resolution. The EyeLink1000 was calibrated using a 9-point calibration. The calibration was monitored and corrected between each trial via an additional drift correction to ensure an optimal calibration.

In order to determine scores for trial accuracy in each version of the Eye-Tracking task, eye-movement data were analyzed offline using BeGaze (mobile eye tracker) and DataViewer (desktop eye tracker) software. Prior to analysis, each participant's data were checked to ensure an accurate calibration; when offset corrections were required, they were implemented by measuring eye gaze against a frequently fixated object of interest (the blue ball). The mobile eye-tracking data represented a participant's eye movements via a cursor overlaid on the video recording. The eye-movement data were analyzed based on a 3000 ms interest period. This began from the moment the seeker emerged from behind the screen (Fig. 7a) and ended when the seeker was positioned behind the table (Fig. 7b). Data were manually coded to indicate the location of the participant's gaze based on eight areas of interest (AOI). These were: the four boxes, the seeker's face, the seeker's body; the hider's face and the hider's body (Fig. 2).

3.4. Procedure

Both the Eye-Tracking task and the Direct Response task featured 12 trials in total: six false-belief experimental trials and six true-belief baseline trials. True-belief trials were included, first to discourage the use of rule-based strategies (e.g., always indicating the box that contained the ball), and second to act as a test of participants' understanding of the task. This meant that any poor performance on the false-belief trials could be ascribed to difficulties with mentalizing (i.e., tracking the seeker's knowledge), rather than to any misunderstanding of the task.

The two conditions (Live and Recorded) followed an identical procedure. Task instructions were given directly by an experimenter for each condition, to rule out the possibility that

improvements in task performance in the Live condition could be due to participant inter-action with the experimenter. In each condition, participants first watched four practice tri-als, then completed the Eye-Tracking task, and then completed the Direct Response task. The Eye-Tracking task practice trials were structured to introduce the participant to the task, introducing a new concept with each phase. The concepts were not verbally stated, but were introduced as part of each role-play: first, the seeker will always search for the ball; second, the hider may lift and replace the boxes when the seeker leaves; third, the hider may swap the boxes when the seeker leaves; and fourth, the seeker would sometimes peek back and observe the boxes being swapped. During these practice trials, the seeker actively searched for the ball, emphasizing her role for the following trials. For the Eye-Tracking task, partici-pants completed six experimental trials presented in a random order (three true-belief trials, where the seeker knew where the ball was, and three false-belief trials, where the seeker held a false belief about where the ball was). They were then offered a break, and then completed another six experimental trials (three true-belief trials and three false-belief trials). During the experimental trials, the seeker did not actually search for the ball; instead, the trial ended just before the search phase would have begun. For the Direct Response task, participants completed three practice trials. Participants were asked "Where will the seeker search for the ball?" These trials could be either true-belief or false-belief trials, and were included to allow the participant to practice the button-press/finger-point response. They then completed six experimental trials, before being offered a break. They then completed another six exper-imental trials.

As participants did not receive feedback on where the seeker would search during the exper-imental trials, they were, therefore, unaware whether their responses on each trial were cor-rect or incorrect. The procedure for the mentalizing tasks was then repeated for the remaining condition. Therefore, in total, each participant completed 48 experimental trials (6 × Live-Eye-Tracking false-belief; 6 × Live-Eye-Tracking true-belief; 6 × Live-Direct-Response false-belief; 6 × Live-Direct-Response true-belief; 6 × Recorded-Eye-Tracking false-belief; 6 × Recorded-Eye-Tracking true-belief; 6 × Recorded-Direct-Response false-belief; and 6 × Recorded-Direct-Response true-belief). Finally, participants were guided through the fun-neled debriefing procedure to check that they had not engaged in explicit mentalizing behavior during the task that measured implicit mentalizing.

4. Results and discussion

Of the 31 participants tested, eight were excluded from the final analysis: two for failure to complete the task, and six for indicating that they had consciously engaged in strategies to process the actor's mindset in the implicit task, as indicated by responses to the funneled debrief. This left a sample of 23 participants (12 females and 11 males), with a mean age of 22 years (range = 19–25, $SD = 2.65$). Due to the use of a within-subjects design, the order of presentation of the Live and Recorded conditions was counterbalanced between participants; preliminary analyses confirmed that there were no order effects present within the data as a result of this manipulation.[1]

4.1. First fixation and direct response analysis

The critical analyses focused on whether participants performed differently on each task depending on whether it was completed in the Live or Recorded condition. Statistical analyses aimed to investigate whether there were differences on the Eye-Tracking Task between the Live and Recorded conditions for the AOI first fixated. The analyses also aimed to investigate if there were differences on the Direct Response task between the Live and Recorded conditions.

A 2×2 repeated measures ANOVA with two factors of condition (Live/Recorded) and task type (Eye-Tracking/Direct Response) was conducted on the average number of correct responses. This revealed a main effect of condition ($F(1,22) = 10.45$, $p = .023$, $\eta\rho^2 = 0.21$), as the average number of correct responses was greater in the Live condition, and a main effect of task type ($F(1,22) = 48.18$, $p < .001$, $\eta\rho^2 = 0.69$), as the average number of correct responses was greater for the Direct Response task. There was no interaction between condition and task, ($F(1,22) = 2.03$, $p = .168$, $\eta\rho^2 = 0.08$) indicating that the improvement in task performance in the Live condition did not differ between the two different task types. This demonstrates that the physical presence of other people can influence a participant's responses while completing mentalizing tasks.

4.2. Preferential looking score and direct response analysis

To further investigate differences in eye-movement behavior between the Live and Recorded conditions for the Eye-Tracking task, a preferential looking score was calculated for each participant, to determine which box the participants preferentially looked toward over the course of the experimental trials. Preferentially attending to the box congruent with the seeker's false belief was considered to be the "correct" response. To allow comparison with the Direct Response task, the proportion of correct responses a participant had given was calculated for the Direct Response task.

A 2×2 repeated measures ANOVA with two factors of condition (Live/Recorded) and task type (Eye-Tracking/Direct Response) was conducted on the average number of correct responses. This revealed that there was no main effect of condition ($F(1,22) = 3.69$, $p = .068$, $\eta\rho^2 = 0.14$), but that there was a main effect of task type ($F(1,22) = 26.90$, $p < .001$, $\eta\rho^2 = 0.55$), as the average number of correct responses was greater for the Direct Response task. There was no interaction between condition and task, ($F(1,22) = 0.61$, $p = .443$, $\eta\rho^2 = 0.03$), indicating that the better task performance shown for the Direct Response task was not influenced by the condition either task was completed in. The results of this analysis do not mirror those of the analysis on the first fixation data; however, there was a trend for participants to preferentially attend more to the correct box in the Live condition (Mean proportion correct $= 0.61$) than in the Recorded condition (Mean proportion correct $= 0.47$).

4.3. Planned comparisons

In addition to the standard tests, the following sections also present the results of Bayesian analyses to indicate the strength of the effects found. In these sections, BF_{H1} indicates values
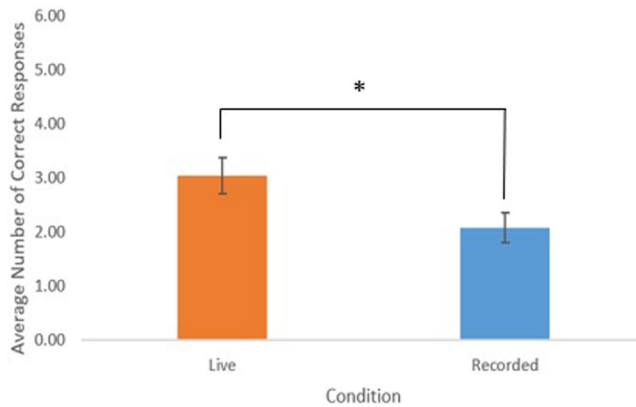
Fig. 3. Average number of Eye-Tracking-task correct first fixation responses in the Live and Recorded conditions. Error bars show $+/-1$ within-subject standard error of the mean (S.E.M.).[2]

in favor of the alternate hypothesis, and BF_{H0} indicates values in favor of the null hypothesis. As proposed by Lee and Wagenmakers (2013), a Bayes factor of greater than 3 is taken as moderate evidence for the relevant hypothesis.

4.3.1. First fixation analysis

Planned comparisons aimed to investigate how each individual task (Direct or Eye-Tracking) was influenced by a social presence. First, statistical analyses aimed to investigate whether there were differences between the Live and Recorded conditions for the AOI first fixated. A Bonferroni-corrected paired samples t-test revealed a significant difference in accuracy between the conditions, $t(22) = 3.14$, $p = .005$, $d = 0.65$ ($BF_{H1} = 9.20$; $BF_{H0} = 0.01$). The results revealed that participants were more accurate on the Eye-Tracking task when they completed it in the Live condition than when they completed it in the Recorded condition (Fig. 3).[2]

4.3.2. Preferential looking score

Further planned comparisons then investigated whether this result was also apparent for the preferential looking score data. A Bonferroni-corrected paired samples t-test revealed that there was no significant difference between the Live and Recorded conditions in the proportion of time spent looking at the correct box ($t(22) = 1.77$, $p = .091$, $d = 0.37$ ($BF_{H1} = 0.83$; $BF_{H0} = 1.21$)); however, there was a trend for participants to preferentially attend more to the correct box in the Live condition (Mean proportion correct $= 0.61$) than in the Recorded condition (Mean proportion correct $= 0.47$) (Fig. 4).[3]

4.3.3. Explicit mentalizing

It was anticipated that participants would generally perform well on the Direct Response task, with the potential that the data would reveal a ceiling effect. However, for completeness, a comparison of the responses from the Live and Recorded conditions was also conducted.
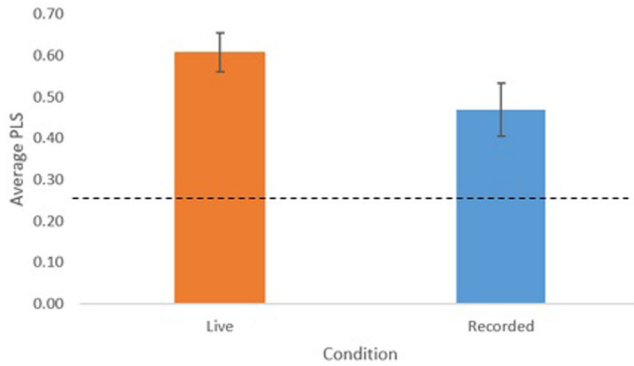
Fig. 4. Average preferential looking score in the Live and Recorded conditions. Error bars show $+/-1$ within-subject standard error of the mean (S.E.M.).[3]
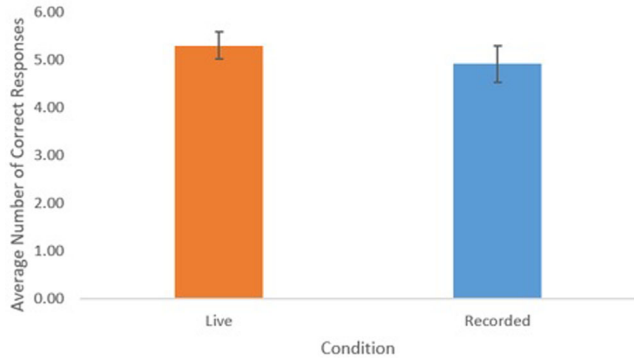

Fig. 5. Average number of Direct Response task correct responses in the Live and Recorded conditions. Error bars show $+/-1$ within-subject standard error of the mean (S.E.M.).

Data for the Direct Response task were not normally distributed, therefore, a Bonferroni-corrected Wilcoxon signed-rank test was used to investigate whether participants performed differently on the Direct Response tasks in the Live and Recorded conditions. There was no difference in accuracy between the conditions, $Z = -0.823$, $p = .410$, $r = .172$ ($BF_{H1} = 0.36$; $BF_{H0} = 2.78$). This indicates that participants did not perform differently on the Direct Response task in the Live and Recorded conditions. However, it was not possible to draw definitive conclusions as to the effect of a social presence on explicit mentalizing, due to the participants' high level of accuracy on the Direct Response task, which may have led to a ceiling effect which influenced the outcomes of the ANOVA analyses (Fig. 5).

4.4. Implicit mentalizing error analysis

4.4.1. First fixation errors

To investigate whether the participants' levels of task engagement were similar across each condition, analyses were conducted on gaze behavior when the participants' first fixation was

Table 1

Percentage of first fixations to each AOI in the Eye-Tracking task

Area of interest	Live (%)	Recorded (%)
Correct box	54.35	44.20
Box containing the ball	7.97	10.14
Other box	1.45	4.35
Hider	2.17	7.97
Background	12.32	10.87
Seeker	21.74	22.46

Table 2

Percentage of participants preferentially attending to each AOI for the Eye-Tracking task

Area of interest	Live (%)	Recorded (%)
Correct box	59.42	37.68
Box containing the ball	16.67	28.99
Other box	0.72	0.72
Hider	0.00	3.62
Background	6.52	6.52
Seeker	16.67	21.01

not consistent with belief tracking. The locations of these "incorrect" fixations were analyzed to determine where the participants directed their gaze instead. Fixations were coded based on the AOIs indicated in Fig. 4. The percentage of first fixations to each AOI is listed in Table 1.

It can be seen from the data presented in Table 1 that the fixations made in each condition were broadly similar, with the three most commonly fixated AOIs on "incorrect" trials being the seeker, the background, and the box containing the ball. Bonferroni-corrected Wilcoxon signed-rank tests were performed to check for differences in the types of errors made between the Live and Recorded conditions. The only significant difference found between conditions was that participants would look to the hider significantly less often in the Live condition compared to the Recorded condition ($Z = 2.83, p = .005, r = .59$). The results of this analysis, therefore, suggest that participants displayed similar levels of attention to each AOI in both conditions.

4.4.2. Preferential looking errors

To further investigate whether participants' levels of task engagement were similar across each condition, analyses were conducted on the participants' gaze behavior when participants preferentially attended to a location not consistent with belief tracking. The preferential looking data for trials coded as "incorrect" were inspected, and the AOI with the greatest gaze duration was noted. The trials where each AOI category had been preferentially gazed at were then summed and the percentage of participants preferentially attending to each AOI was calculated (Table 2).

Consistent with the first fixation error analysis, it was found that similar errors were made in each condition, with the three most commonly fixated AOIs being the box containing the ball, the seeker, and the background. Bonferroni-corrected Wilcoxon signed-rank tests confirmed that there were no significant differences between the Live and Recorded conditions for which AOI was preferentially attended to. Again, this confirms that participants found both conditions equally engaging, with errors arising from comparable sources across both contexts. In combination with the first fixation error analysis, it is clear that participants demonstrated a preference to attend both to the box containing the ball and to the seeker. Participants were, therefore, more likely to attend to the AOIs that were directly related to the narrative of each trial, rather than to randomly distribute their attention to alternate areas of the scene. This suggests that participants' patterns of looking behavior were not influenced by visual field differences between the methods of presentation used in each condition (computer screen vs. real actors). Further, this demonstrates that participants continued to show task-contingent eye-movement behavior even on incorrect trials and, consequently, that when errors occurred, they were not related to a lack of task engagement in either condition.

5. Study 2

Out of the 23 participants tested in study 1, 19 scored at ceiling level on the Direct Response task in the Live condition, and 17 scored at ceiling level in the Recorded condition. While this demonstrates that participants were able to clearly comprehend and engage with the task, the very narrow range of scores prevents the identification of any potential differences in task performance between the Live and Recorded conditions. Therefore, a second study was conducted, using a more challenging mentalizing task.

Study 2 used a nonverbal second-order theory of mind task (adapted from Apperly, Samson, Carroll, Hussain, & Humphreys, 2006). A second-order task requires participants to understand that a protagonist holds a false belief about another person's belief. As in study 1, there was both a Direct Response and Eye-Tracking version of the task, and each task was completed in a Live condition (where task protagonists were physically present) and a Recorded condition (where the same task protagonists were depicted in a video recording).

Study 2 aimed to further investigate the effect of social presence on explicit and implicit mentalizing. Following the findings of study 1, it was expected that the participants in study 2 would demonstrate significantly improved task performance when completing the tasks in the Live condition compared to the Recorded condition. Based on the previous research (Chevallier et al., 2014; Foulsham, Walker, & Kingstone, 2011; Freeth, Foulsham, & Kingstone, 2013; Laidlaw, Foulsham, Kuhn, & Kingstone, 2011), it was, therefore, predicted that performance on both the Eye-Tracking and Direct Response tasks would be enhanced by a social presence in the Live condition, compared to the Recorded condition of the task. If participants displayed different patterns of behavior in the Live compared to the Recorded condition—for either the Eye-Tracking or Direct Response task—then, this would suggest that there are fundamentally different cognitive mechanisms engaged when real people are physically present.

6. Method

6.1. Participants

Based on a power calculation conducted using the observed effect from study 1, study 2 recruited 50 adults (24 females and 26 males) via opportunity sampling through a university-wide volunteers list. All participants received a gift voucher as a thank you for taking part.

Of the 50 participants tested, five were excluded from the final analysis: one participant was excluded due to failure to complete the task; two were excluded due to poor eye-tracking calibrations; one participant scored fewer than 6 out of 12 correct responses on the direct true-belief trials; and one participant was excluded for misunderstanding the task instructions. This left a final sample of 45 participants (22 females and 23 males), with a mean age of 35 years (range $= 18$–66, $SD = 13.19$).

6.2. Design

Study 2 used an identical design to study 1, namely, a within-subjects design with two independent variables: condition (Live or Recorded) and task type (Eye-Tracking or Direct Response). The key difference from study 1 was that instead of a first-order task, a nonverbal second-order theory of mind task was used. This task required participants to track the knowledge state of three protagonists.

The introduction of a third protagonist altered the procedure of the role-play from that used in study 1. During the second-order task, the participant watched as three protagonists (the "hider," the "seeker," and the "helper") acted out a role-play in which the hider hid a ball, the seeker tried to find the ball, and the helper tried to assist with the seeker's search (Fig. 6). The participant was told that the helper would always try to help the seeker (by placing a small red marker on the box they believed contained the ball).

Each trial would begin with the helper watching the hider hide a ball in one of four boxes; at this time, the seeker was not present in the scene, and, therefore, did not witness the hiding of the ball. After the ball was hidden, the seeker would enter the scene and join the helper and hider; the seeker, therefore, "knew" that the helper had seen where the ball was hidden. The helper would then leave the scene, and the hider would manipulate two visually identical boxes—either swapping their positions, or picking them up and replacing them in the same positions. On a number of trials, the helper would peek back and watch the scene, meaning that on those trials, they could accurately track the ball's location. Participants were told in advance that the seeker would not know whether the helper peeked on each trial. After the hider had manipulated the boxes, the helper would then re-enter the scene and place the red marker on the box they believed contained the ball. The seeker would then step forward as if to search for the ball.

In order to be able to predict where the seeker would search for the ball, participants needed to understand what the seeker believed about what the hider believed. That is, participants needed to understand that the seeker did not know if the hider had peeked. For the Eye-Tracking task, participants were instructed simply to watch the role-plays; during these tasks, participants' eye movements were recorded in order to measure whether participants
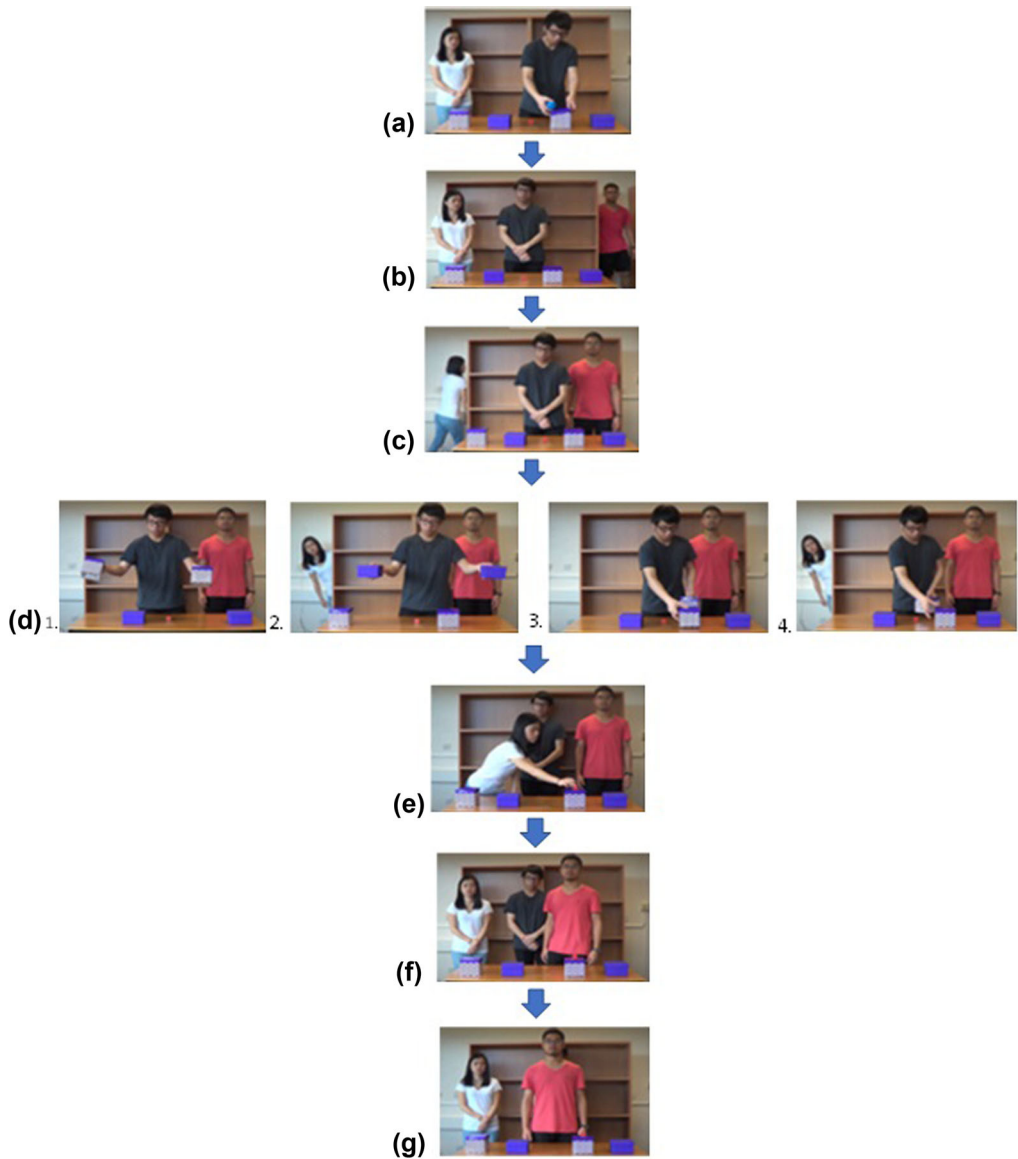
Fig. 6. The experimental procedure. The trial type is labeled based on the seeker's belief. (a) The hider hides the ball, watched by the helper. (b) The seeker appears; he is aware that the ball has been hidden and that the helper knows where it is. (c) The helper leaves. (d) (1) True-belief trial: the hider lifts the boxes and replaces them in their original positions; the helper remains hidden. (2) True-belief trial: the hider lifts the boxes and replaces them in their original positions; the helper watches; the seeker is not aware the helper has peeked. (3) Second-order true-belief trial: the hider swaps the boxes, the helper remains hidden. (4) Second-order false-belief trial: the hider swaps the boxes, the helper watches; the seeker is not aware the helper has peeked. (e) The helper places the clue (red cube). (f) The seeker steps forward to search for the ball. (g) The seeker remains standing behind the table. For the Direct Response trials, the participant provides a response at this point.
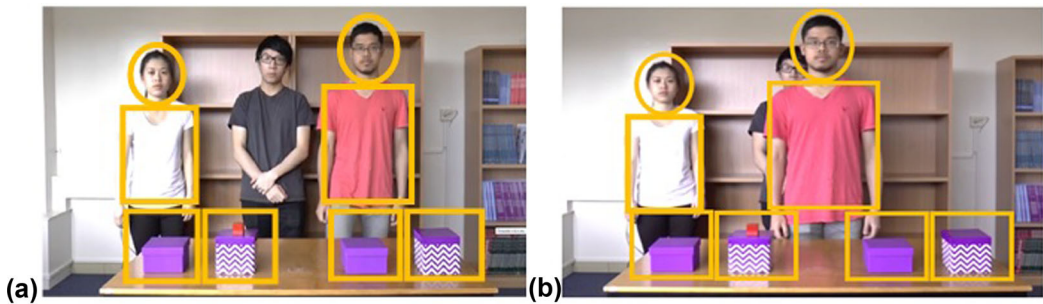
Fig. 7. The areas of interest for the anticipatory gaze data analysis. (a) At the start of the interest period, as the seeker steps forward from the bookcase. (b) At the end of interest period, while the seeker remains behind the table.

spontaneously tracked the seeker's false belief. For the Direct Response task, participants were asked to indicate which box they thought the seeker would look in, either by making a finger point (for the Live condition) or by a button press (for the Recorded condition).

6.3. Apparatus and procedure

Study 2 used the same apparatus and procedure as detailed in study 1. However, the AOI used to code the eye-tracking data were updated to reflect the change in task. In study 2, the data were manually coded to indicate the location of the participant's gaze, based on eight AOI. These were: the four boxes, the seeker's face, the seeker's body; the helper's face and the helper's body (Fig. 7).

7. Results

7.1. First fixation and direct response analysis

Statistical analyses aimed to investigate whether there were differences on the Eye-Tracking Task between the Live and Recorded conditions for the AOI first fixated. The analyses also aimed to investigate if there were differences on the Direct Response task between the Live and Recorded conditions. A 2×2 repeated measures ANOVA with two factors of condition (Live/Recorded) and task type (Eye-Tracking/Direct Response) was conducted on the average number of correct responses. This revealed a main effect of condition ($F(1,44) = 5.57$, $p = .023$, $\eta\rho^2 = 0.11$), as the average number of correct responses was greater in the Live condition than in the Recorded condition, and a main effect of task type ($F(1,44) = 24.48$, $p < .001$, $\eta\rho^2 = 0.36$), as the average number of correct responses was greater for the Direct Response task than for the Eye-Tracking task. There was also an interaction between condition and task, ($F(1,44) = 14.16$, $p < .001$, $\eta\rho^2 = 0.24$), indicating that the improvement in task performance in the Live condition differed between the two different task types (Eye-Tracking vs. Direct Response).
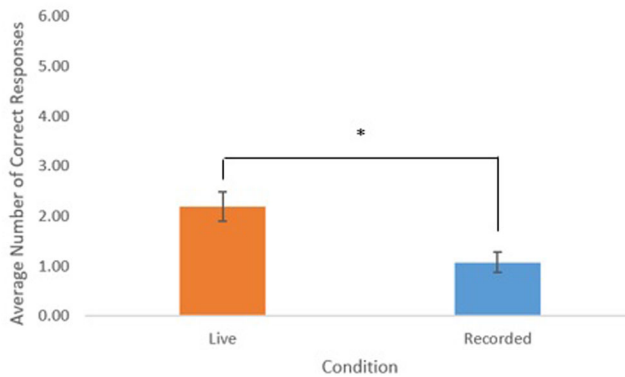
Fig. 8. Average number of Eye-Tracking-task correct first fixation responses in the Live and Recorded conditions. Error bars show $+/-1$ within subject standard error of the mean (S.E.M.).[4]
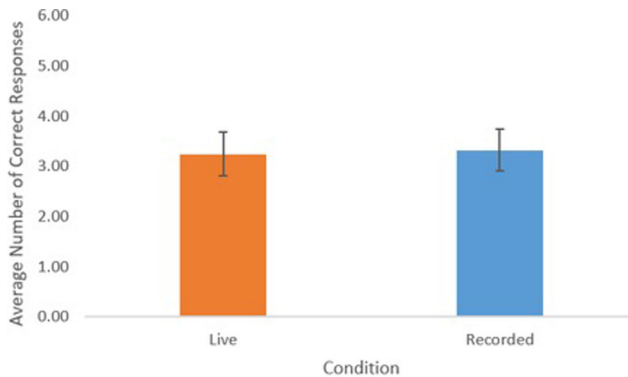

Fig. 9. Average number of Direct Response-task correct responses in the Live and Recorded conditions in study 2. Error bars show $+/-1$ within subject standard error of the mean (S.E.M.).

In order to investigate the interaction between condition and task, two separate post hoc tests were conducted, one for the Eye-Tracking Task and one for the Direct Response Task. In addition to the standard tests, the following sections also present the results of Bayesian analyses to indicate the strength of the effects found. As proposed by Lee and Wagenmakers (2013), a Bayes factor of greater than 10 is taken as strong evidence for the alternate hypothesis. The first analysis investigated whether the AOI that participants first fixated differed depending on whether the task was completed in the Live or Recorded condition. A Shapiro–Wilk test for normality showed that the data were not normally distributed ($p < .05$); a Bonferroni-corrected Wilcoxon signed-rank test revealed that participants were significantly more accurate on the Eye-Tracking task in the Live condition than in the Recorded condition, $Z = -3.26$, $p = .001$, $r = .49$ ($BF_{H1} = 49.40$; $BF_{H0} = 0.02$) (Fig. 8).[4]

However, participants did not show the same improvement between the Recorded and Live conditions on the Direct Response Task (Fig. 9). A Shapiro–Wilk test for normality showed that the data were not normally distributed ($p < .05$), therefore, a Wilcoxon signed-rank test

was run to determine whether there was a difference in the number of accurate responses given in the Live and Recorded conditions. Participants were not significantly more accurate on the Direct Response task in the Live compared to the Recorded condition, $Z = 0.29$, $p = .775$, $r = .04$ ($BF_{H1} = 0.16$; $BF_{H0} = 6.17$). This differs from the findings of the Eye-Tracking task, where the measure of implicit mentalizing was clearly affected (and facilitated) by the task being conducted live rather than being presented via a recording.

7.2. Preferential looking score and direct response analysis

To further investigate differences in eye-movement behavior between the Live and Recorded conditions for the Eye-Tracking task, a preferential looking score was calculated for each participant, to determine which box the participants preferentially looked toward over the course of the experimental trials. Preferentially attending to the box congruent with the seeker's false belief was considered to be the "correct" response. To allow comparability with the Direct Response task, the proportion of correct responses a participant had given was calculated for the Direct Response task.

A 2×2 repeated measures ANOVA with two factors of condition (Live/Recorded) and task type (Eye-Tracking/Direct Response) was conducted on the average number of correct responses. This revealed a main effect of condition ($F(1,44) = 7.50$, $p = .009$, $\eta\rho^2 = 0.15$), as the average number of correct responses was greater in the Live condition than in the Recorded condition; and a main effect of task type ($F(1,44) = 25.79$, $p < .001$, $\eta\rho^2 = 0.37$), as the average number of correct responses was greater for the Direct Response task than for the Eye-Tracking task. There was also an interaction between condition and task, ($F(1,44) = 15.28$, $p < .001$, $\eta\rho^2 = 0.26$), indicating that the improvement in task performance in the Live condition differed between the two different task types (Eye-Tracking vs. Direct Response).

In order to investigate the interaction between condition and task, post hoc tests were conducted for the Eye-Tracking Task. (Follow-up analyses for the Direct Task are reported in Section 7.1: the "First Fixation" analysis.) Bayesian analyses were also conducted, and are reported below. A Bonferroni-corrected Wilcoxon signed-rank test revealed a significant difference in preferential looking behavior between the Live and Recorded conditions ($Z = -3.71$, $p < .001$, $r = .55$ ($BF_{H1} = 301.09$; $BF_{H0} = 0.003$)), with participants preferentially attending to the correct box more in the Live condition than in the Recorded condition (Fig. 10). Therefore, participants were significantly more accurate at the Eye-Tracking task when it was completed with a social presence, as confirmed by the participants' first fixation data and preferential looking score. This confirms that the measure of implicit mentalizing was clearly affected by the task being conducted live rather than being presented via a recording, and that such measures used in mentalizing tasks are susceptible to influence by a social presence.

7.3. Implicit data error analysis

7.3.1. Implicit mentalizing first fixation error analysis

To investigate whether participants' levels of task engagement were similar across each condition, analyses were conducted on gaze behavior when the participants' first fixation was
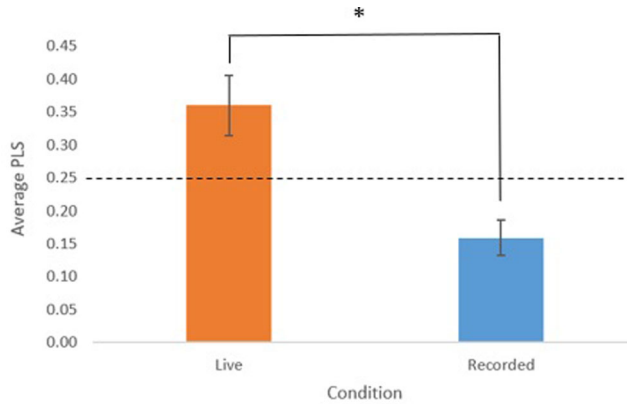
Fig. 10. Average preferential looking score in the Live and Recorded conditions.[5] Error bars show $+/-1$ within subject standard error of the mean (S.E.M.).

Table 3
Percentage of participants who first fixated each AOI on the incorrect trials for the Eye-Tracking task in the Live and Recorded conditions in study 2

Area of interest	Live (%)	Recorded (%)
Correct box	35.56	18.43
Box containing the ball	27.41	35.29
Other boxes	3.33	10.2
Helper	2.59	5.49
Hider	0.00	0.39
Background	4.44	8.63
Seeker	26.67	21.57

not consistent with belief tracking. The eye-movement data from the incorrect trials for the Eye-Tracking task were analyzed (Table 3). Errors were analyzed based on the AOIs indicated in Fig. 6. Data were analyzed using Bonferroni-corrected Wilcoxon signed-rank tests to check for differences in the types of errors made between the Live and Recorded conditions. Of note, the analysis revealed that participants were significantly more likely to first fixate one of the "other boxes," that is, boxes that had never contained the ball, in the Recorded condition than in the Live condition ($Z = -2.81$, $p = .005$, $r = .42$). This analysis suggests that participants followed the narrative of the trial less well in the Recorded condition. The "other boxes" were not directly relevant to the narrative of each trial, and did not play an active role in the role-play acted out by the hider and the seeker. In the Live condition, participants were more likely to successfully track the narrative of each trial and attend to the two boxes relevant to each role-play. This suggests that social presence influenced looking behavior, leading to changes in the types of errors made between each condition.

Table 4
Percentage of participants preferentially attending to each AOI for the error trials for the Eye-Tracking task

Area of interest	Live (%)	Recorded (%)
Correct box	45.70	21.67
Box containing the ball	27.73	32.08
Other boxes	3.13	5.00
Helper	0.79	1.25
Hider	0.00	0.42
Background	1.56	2.08
Seeker	21.09	37.50

7.3.2. Preferential looking errors

To investigate participants' levels of task engagement across conditions, the preferential looking data for trials coded as incorrect were inspected. The proportion of overall looking time directed to each AOI was calculated for each participant (Table 4).

It was found that participants made similar errors across each condition, most commonly attending to the AOIs directly related to the narrative of each trial, that is, either the seeker or the box containing the ball. The data were then further analyzed using Bonferroni-corrected Wilcoxon signed-rank tests, and revealed no significant differences in the types of errors made across each condition.

Importantly, the error analyses from both study 1 and study 2 demonstrate consistently that even on incorrect trials, participants' eye movements tended to align with the AOIs most relevant to the narrative of each trial—with both first fixation and preferential looking errors most likely to occur from attention to either the box containing the ball or to the seeker. This suggests that participants' patterns of looking behavior were not influenced by visual field differences between the methods of presentation used in each condition (computer screen vs. real actors). Further, this suggests that participants remained engaged with the task across each condition and, critically, the results of these studies also reveal that participants were more likely to show gaze behavior associated with belief tracking (i.e., attending to the correct box) in the Live condition than in the Recorded condition. This, therefore, has implications for the reliability and comparability of gaze behavior recorded for false belief tasks completed across different contexts. The measure of implicit mentalizing used in these types of tasks has been demonstrated to be sensitive to a social presence (i.e., the mere presence of other people). Therefore, serious consideration should be given to social presence when designing paradigms intended to investigate the existence of an implicit mentalizing system.

8. General discussion

The current study investigated whether a common method of investigating and measuring mentalizing was significantly affected by the presence of others when those "others" were

physically present. The answer was a very clear "yes": it was found that the measure of implicit mentalizing behavior was affected by a social presence, with participants demonstrating significantly different patterns of gaze behavior in response to a social presence, in both a first-order false-belief task (study 1) and a second-order false-belief task (study 2). In contrast, explicit mentalizing was found to be unaffected by a social presence, with no difference in performance according to whether or not real people were present.

There are a range of implications arising from this finding, of which the most important is that research paradigms lacking a social presence may not, after all, serve as suitable analogues for real-world social interactions. This is a critical finding, as it has real consequences for implicit mentalizing research. This paper presents cumulative evidence demonstrating that social presence directly influences how people respond in mentalizing paradigms. Crucially, if participants demonstrate quantifiable behavioral changes in response to a social presence, then social cognition paradigms conducted in social isolation are unlikely to reflect how cognitive processes function in real-world environments. This research, therefore, supports previous studies, which argue that social interaction is inherent to understanding social cognition (De Jaegher et al., 2010; Gobel et al., 2015), by clearly demonstrating that social presence influences measures of mentalizing. Together, these studies highlight the urgent need to use ecologically valid tasks in social cognition research. Further, these findings raise important concerns regarding the generalizability of paradigms that are conducted without a social presence.

In both studies, participants' eye movements were altered by a social presence, with the altered behavior appearing in line with what would be expected with the engagement of implicit mentalizing processes. The results of this paper, therefore, lend support to studies which have provided either partial or full replications of some of the tasks used in implicit mentalizing research (Powell, Hobbs, Bardis, Carey, & Saxe, 2018; Schuwerk, Priewasser, Sodian, & Perner, 2018). It has been argued that rather than not existing at all, implicit mentalizing may instead be a "fragile" construct, which is difficult to capture and measure in a consistent fashion (Kulke et al., 2019; Kulke, von Duhn, Schneider, & Rakoczy, 2018). While the results of the current paper do not provide conclusive evidence for the existence of an implicit mentalizing system, they are consistent with the argument that a potentially important factor in eliciting implicit mentalizing behavior may be the inclusion of naturalistic stimuli. This accords with suggestions from recent research that it may only be through the use of more engaging, naturalistic stimuli that we will be able to capture the empirical evidence of such an unconscious system (Krupenye et al., 2016; Kulke et al., 2018; Kulke et al., 2019).

The results of this paper present formative evidence in a relatively new area of research. However, what remains to be determined are the precise mechanisms that underlie the behavioral changes observed in this study. While it could be suggested that social presence might serve to make tasks more engaging, the data from this paper suggest that this cannot be the sole motivating factor for the improvement in task performance. In the first instance, participants' eye movements on error trials (when participants did not look to the correct box) demonstrated that their gaze remained task-focused and directed to the stimuli relevant to

narrative of each trial, across both the Live and Recorded conditions. Moreover, analyses revealed that there were no significant differences in participants' specific patterns of looking behavior, with similar amounts of looking time and first glances dedicated to each stimulus, regardless of condition. This demonstrates that participants engaged with the task similarly well in each condition, with the sole differences in gaze behavior relating to eye movements associated with the measure of implicit mentalizing. Further, across both the Direct task and Eye-Tracking task, we observed significant improvements on the measure of implicit mentalizing *only*, with no improvement on the measure of explicit mentalizing. This, therefore, suggests that social presence does not solely act to increase motivation, and future research is required to identify the mechanisms underlying these behavioral changes.

These findings also have important implications for mentalizing studies conducted with autistic participants. Previous research has demonstrated that autistic participants consistently show difficulties with implicit mentalizing; however, this research has typically been carried out using video-based stimuli (Senju, Southgate, White, & Frith, 2009; Schneider, Slaughter, Bayliss & Dux, 2013; Schuwerk, Jarvers, Vuori, & Sodian, 2016). Critically, in the present study, neurotypical adults were significantly more accurate on a false belief task when it was completed with a social presence. The finding that paradigms without a social presence lack the sensitivity to elicit behavior associated with implicit mentalizing also raises the possibility that the differences reported on measures of implicit mentalizing for autistic participants may be, wholly or in part, a function of the experimental paradigm used. It would, therefore, be important to investigate whether autistic participants also show different patterns of behavior associated with implicit mentalizing in response to a social presence, and whether their previously reported difficulties generalize to situations where real people are present.

9. Conclusion

The results presented in this paper provide consistent evidence that adults are sensitive to a social presence, and that this sensitivity can lead to quantifiable changes in gaze behavior when completing a false belief task. Importantly, these findings have far-reaching implications for social cognition paradigms lacking a social presence, and suggest that results from such paradigms should not unreflectively be considered analogous to real-life behavior. This study, therefore, highlights the critical importance of understanding the effect of a social presence, and suggests that the inclusion of a social presence needs to be given strong consideration across social cognition paradigms.

10. Data deposition information

The data that support the findings of this study are available in the Open Science Framework at https://osf.io/2vk8c/?view_only=f6b59f8153b0486ca4c54600d4d803d9

Notes

1. Wilcoxon Signed-Rank tests revealed no differences between the first and second iterations of either the first fixation data or the Direct data (First Fixation $p = .412$; Explicit $p = .503$), which was confirmed via Bayesian analyses (First Fixation $BF_{H1} = 0.12$, $BF_{H0} = 8.22$; Direct Task $BF_{H1} = 0.36$, $BF_{H0} = 2.78$). Paired samples t-tests revealed no differences in accuracy scores between the first and second iterations of the implicit preferential looking data ($p = .068$ ($BF_{H1} = 0.08$, $BF_{H0} = 12.14$)). This confirms that completion of the first task did not significantly influence performance on the second task.
2. Participants were significantly more likely to look at the correct box first in both the Live ($p < .001$) and Recorded ($p < .05$) conditions than would be expected if there were no systematic bias to attend to the correct box.
3. Participants looked at the correct box in both the Live ($p < .001$) and Recorded ($p < .05$) conditions for significantly longer than would be expected if there were no systematic preference to attend to the correct box.
4. In the Live condition, participants were significantly more likely to look at the correct box first than would be expected if there was no systematic looking behavior ($p = .030$); conversely, in the Recorded condition, they displayed a systematic bias to attend to one of the incorrect boxes ($p = .021$).

Acknowledgments

Open Research Badges

This article has earned Open Data badges. Data are available at https://osf.io/7vn4f/.

References

Aihara, T., Yamamoto, S., Mori, H., Kushiro, K., & Uehara, S. (2015). Observation of interactive behaviour increases corticospinal excitability in humans: A transcranial magnetic stimulation study. *Brain and Cognition*, *100*, 1–6.

Apperly, I. A., Samson, D., Carroll, N., Hussain, S., & Humphreys, G. W. (2006). Intact 1st and 2nd order false belief reasoning in a patient with severely impaired grammar. *Social Neuroscience*, *1*(3–4), 334–348.

Bargh, J. A., & Chartrand, T. L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–285). Cambridge University Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology*, *4*, 113–125.

Birmingham, E., & Kingstone, A. (2009). Human social attention: A new look at past, present, and future investigations. *Annals of New York Academy of Sciences*, *1156*, 118–140.

Chevallier, C., Parish-Morris, J., Tonge, N., Le, L., Miller, J., & Schultz, R. T. (2014). Susceptibility to the audience effect explains performance gap between children with and without autism in a theory of mind task. *Journal of Experimental Psychology: General*, *143*(3), 972–979.

Cole, G. G., Skarratt, P. A., & Kuhn, G. (2016). Real person interaction in visual attention research. *European Psychologist*, *21*(2), 141–149.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, *14*, 441–447.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920–1931.

Freeth, M., Foulsham, T., & Kingstone, A. (2013). What affects social attention? Social presence, eye contact and autistic traits. *PLoS One*, *8*(1), e53286.

Gawronski, B., & Hahn, A. (2019). Implicit measures: Procedures, use, and interpretation. In H. Blanton, J. M. LaCroix, & G. D. Webster (Eds.), *Measurement in social psychology* (pp. 29–55). New York: Taylor & Francis.

Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, *136*, 359–364.

Gregory, N. J., Lopez, B., Graham, G., Marshman, P., Bate, S., & Kargas, N. (2015). Reduced gaze following and attention to heads when viewing a "Live" social scene. *PLoS One*, *10*(4). https://doi.org/10.1371/journal.pone.0121792

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*, 888–900.

Kulke, L., Wubker, M., & Rakoczy, H. (2019). Is implicit Theory of Mind real but hard to detect? Testing adults with different stimulus materials. *Royal Society Open Science*, *6*. https://doi.org/10.1098/rsos.190068

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*, 110–114.

Laidlaw, K. E. W., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Science*, *108*, 5548–5553.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*, 40–50.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioural and Brain Sciences*, *4*(4), 515–629.

Reader, A. T., & Holmes, N. P. (2016). Examining ecological validity in social interaction: Problems of visual fidelity, gaze, and social potential. *Culture and Brain*, *4*, 134–146.

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews: Neuroscience*, *20*, 495–505.

Risko, E. F., Laidlaw, K., Freeth, M., Foulsham, T., & Kingstone, A. (2012). Social attention with real vs reel stimuli: Toward an empirical approach to concerns about ecological validity. *Frontiers in Human Neuroscience*, *6*, 143.

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science*, *25*(1), 70–74.

Rosenblau, G., Kliemann, D., Heekeren, H. R., & Dziobek, I. (2015). Approximating implicit and explicit mentalizing with two naturalistic video-based tasks in typical development and autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *45*(4), 953–965.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2011). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*, 433–438.

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*, 410–417.

Schulze, C., & Tomasello, M. (2015). 18-month-olds comprehend eye-tracking communicative acts. *Cognition*, *136*, 91–98.

Schuwerk, T., Jarvers, I., Vuori, M., & Sodian, B. (2016). Implicit mentalising persists beyond early child-hood and is profoundly impaired in children with autism spectrum condition. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01696

Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Royal Society Open Science*, *5*. https://doi.org/10.1098/rsos.172273

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*, 883–885.

Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Science*, *105*(2), 809–813.

Van der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track other's beliefs? Evidence from a continuous measure. *Cognition*, *130*, 128–133.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Appendix A

Funneled debriefing procedure (adapted from Schneider et al., 2013)

1. What do you think the purpose of the experiment was?
2. What do you think this experiment was trying to study?
3. Did you think that any of the tasks you did were related in any way?

(If "yes") In what way were they related?

4. Did anything you did on one task affect what you did on any other task?

(If "yes") How exactly did it affect you?

5. Did you notice anything unusual about the movies/role-plays?
6. Did you notice any particular patterns or themes?
7. What were you trying to do while watching the movies/role-plays? Did you have any particular goal or strategy?
8. If thinking about the four boxes on the desk, which box do you think you spent the most time on?
10. Did you notice that the actor sometimes had a true belief about the ball location and sometimes had a false belief about the ball location when coming back into the room?

(If participant is unsure ask: Did you notice that the actor was sometimes tricked about the ball location when coming back into the room?)

(If "yes") How did those beliefs become true or false?