Spring 5-31-2023

What you don't know matters: An ignorance-focused investigation of theory of mind

Steven M. Shin
steven.m.shin.23@dartmouth.edu

Jonathan S. Phillips
Dartmouth College, Jonathan.S.Phillips@Dartmouth.edu

What you don't know matters:

An ignorance-focused investigation of theory of mind

Steven Shin

Advised by: Jonathan Phillips

Honors Thesis

Department of Cognitive Science

Dartmouth College

Hanover, NH

Abstract

This project examines the ways in which knowledge, or ignorance, impact healthy adults' theory of mind (i.e. their considerations of others' mental states). In a pilot study, and four experiments, an effect is found which supports the hypothesis that knowledge states influence the execution of theory of mind. The present findings suggest that attention is directed differently when participants reason from positions of knowledge, or positions of ignorance, in regard to a task-relevant fact. This project provides a starting point for further research, investigating the rich contextual contributions to the fluent functioning of the 'theory of mind system.'

Table of Contents

Introduction

Theory of Mind

'Theory of mind' (ToM) is the cognitive capacity that allows us to recognize and reason about the minds of others. When we learn, this ability allows us to identify individuals around us with privileged knowledge. When we cooperate, our ToM helps us to establish common understanding, and to recognize when our cooperators aren't on the 'same page'. And when we lie, or deceive, it is our capacity of ToM which allows us to consider what the deceived knows, what they don't, and what they would be *led to believe* by possible deceptive actions. Theory of mind is integral to every facet of our lives as social creatures.

There is quite a lot that we, as cognitive scientists, have learned about ToM. We know that it's something done fluently by healthy and competent adults. We've identified regions in the brain engaged by ToM (Carrington and Bailey, 2009; Frith and Frith, 2003). And we've made the surprising discovery that linguistically competent children, under the age of four, are unable to perform certain seemingly basic tasks which rely upon ToM (Wimmer and Perner, 1983; Baron-Cohen et al., 1985). But for all that we know, when it comes to many basic aspects of ToM, we're completely in the dark: Is ToM a biologically entailed human ability, or a linguistically-dependent product of cultural evolution (Heyes, 2018)? Can ToM operate implicitly, or does it demand participation from the 'central executive' (Samson, 2008; Apperly and Butterfill, 2009; Heyes, 2014)? Do children under the age of four have ToM at all (Liszkowski et al. 2006, Kovacs et al. 2010; Onishi and Baillargeon, 2005; Doherty, 2008)? Do animals (Call and Santos, 2012; Drayton and Santos, 2016)? What, for that matter, should count as ToM in the first place (Phillips and Norby, 2019)? For all that we know about ToM, it's clear, I

think, that we have quite a long way to go in our understanding of this characteristically human ability. And as we learn more, and better understand the way we think about our own minds and the minds of others, our clarified understanding of ToM will allow us to see farther into the psychological facts that bring us together, tear us apart, and make us special (Sally and Hill, 2006; Paal and Perekzkei, 2007).

Factive and Non-Factive ToM

Before discussing the present and future of ToM research, it is perhaps wise to pay tribute to some historically important developments within the field. In particular, it is important to note that the history of the ToM literature has led the field to focus its efforts on a particular subset of our broader capacity for ToM, known as *non-factive* ToM, while largely neglecting the complementary study of *factive* ToM. Of course, this one-sided approach was not intended. In fact, the *factivity* distinction within discussions of ToM has arisen only recently, as a *response* to the seemingly narrow focus of researchers, not as the motivation for this focus (Phillips and Norby, 2019, Phillips et al. 2021). But this distinction has, nonetheless, had far-reaching impacts upon the literature.

But what exactly is factivity? In precise terms, factive ToM concerns the attribution of mental contents which are consistent with that which is taken to be true, while non-factive ToM concerns the attribution of mental states which are not consistent with that which is taken to be true. It is useful to note that factive ToM generally concerns representations of *knowledge,* while non-factive ToM generally concerns representations of *belief.* To see how this distinction matters, consider a scenario in which Bart, a lovable but dim young rascal, is attempting to play a trick on

his younger sister Lisa. While the two siblings are sitting at breakfast, Bart deposits a fake spider into Lisa's orange juice. We can say, "Bart *knows* that there's a spider in Lisa's orange juice". This knowledge ascription is *factive,* because it's congruent with the way that we (the people making the knowledge ascription) understand the facts. One might also say "he *believes* that a spider is in Lisa's orange juice", but we'll return to that later. Now if Lisa, being quite a bit more observant than Bart, were to take notice of Bart's mischief, and to switch her orange juice with Bart's without his knowledge, the picture would change slightly. Bart would still *believe* that there was a spider in Lisa's orange juice. But given that the spider is, in fact, now in Bart's own glass, it would seem strange to say that "Bart *knows* that there is a spider in Lisa's glass." You can't *know* things that aren't true, you can only *believe* them. Put into the jargon of factivity, knowledge ascriptions are always factive, while belief ascriptions can be factive or non-factive. So while *belief* can be used to discuss Bart's mental states all the while, whether the claims being made are factive or non-factive, *knowledge* ascriptions are only appropriate for *factive* attributions.

The literature has largely focused on non-factive, rather than factive ToM, out of a desire to reduce confounding explanations within ToM tasks. More specifically, researchers have used 'false-belief' tasks, which test non-factive ToM as the gold standard for assessing the capacity of ToM (Baron-Cohen et al, 1985; Phillips and Norby 2019). In a false-belief task, ToM is assessed by testing whether an experimental subject is capable of understanding that an individual can believe something, while the subject knows that the individual's belief is false. This test is useful because proficiency in a false-belief task can only be explained by a capacity for theory of mind (Bennett, 1978; Dennett, 1978; Harman; 1978). For example, in a false-belief task, an experimental subject may be asked to predict where Bart would look for the spider, after Lisa

had covertly changed its location. If the subject is able to correctly predict that Bart would look in Lisa's glass (because Bart was unaware of the switch), they must be able to represent that "Bart *believes* a spider is in Lisa's glass", despite the fact that they know that "the spider is actually in Bart's glass." This is a very strict test of ToM, because it strictly dissociates the subject's own knowledge with the beliefs of the agent in question, necessitating a capacity for ToM. On the other hand, true-belief tasks are conducive to alternative explanations. If a subject is asked to predict where Lisa would look for the spider, they might be able to succeed in their prediction by knowing only that 1) the spider is in Bart's glass and 2) people look for things where they are.

True-belief tasks are thus difficult to interpret because they often fail to demonstrate that a subject's attributed mental states are separate from their own knowledge. Stemming from this concern, many researchers have taken false-belief tasks as the only conclusive evidence for a capacity of ToM. The problem with this approach is that there are lots of instances of legitimate ToM for which non-factive representations are unnecessary (Phillips and Norby, 2019). The present thesis work will seek to broaden our understanding of ToM by examining previously unstudied cases of factive ToM. In particular, this thesis will examine the effects of *egocentric knowledge and ignorance* upon factive ToM attributions.

Egocentric Knowledge and Ignorance

When discussing mental states, we can distinguish between those which are *altercentric* and those which are *egocentric.* Altercentric representations are 'centered on the other.' In other words, they concern the mental states of other people, for example, that "Lisa knows where the

spider is." On the other hand, *egocentric* representations concern one's own mental states, for example, that "I know where the spider is." Throughout this project, 'egocentric' will be used to refer to the mental states of the subjects being studied (the experimental participants) while 'altercentric' will be used to refer to the mental states of other people, about whom the experimental subjects are making considerations. In other words, 'the self' will be identified with the source of the behavioral data in consideration— 'the self' according to the perspectives we are studying. It is clear that each of us is capable of both altercentric or egocentric attribution of mental states. This latter attribution of egocentric mental states is generally factive, as one's own beliefs are usually (if not necessarily) congruent with one's own understanding of the facts. While altercentric mental states are generally the emphasis of ToM research, egocentric mental states (i.e. our own knowledge and beliefs) may also be important to how we consider the mental states of others.

Consider, for example, competitive games like poker which rely heavily on theory of mind. While playing, one is constantly considering the mental states of others. One might infer, from their opponent's sudden smile, that they are pleased with their hand. But these altercentric mental state attributions are closely informed by one's own knowledge states. One is only interested in their opponent's telling smile because 1) they do not know their opponent's hand, 2) their opponent does, and 3) that information is relevant. In a competitive context like a poker game, players are keenly interested in 'filling in the gaps' in their knowledge, understanding that information is most advantageous when it pertains to facts in their environment to which they are *ignorant*. In order to maximize the information gained from every sudden smile or furrowed brow, a poker player must selectively allocate their attention to the aspects of their environment

best suited to filling in these gaps. In order to know where to look, they'll need to be guided by consideration of their own *egocentric knowledge and ignorance.*

In this way, the present project is motivated by a consideration of the kinds of activities and environments which make factive ToM a valuable skill in the first place. One of the things that our ToM allows us to do is to learn things from others in our environment (Phillips et al., 2021). In such cases, egocentric knowledge states are vital to the competent execution of altercentric ToM. In particular, individuals' background knowledge will inform whether, and how, they consider the mental states of others. The present work will investigate the effects of egocentric knowledge states upon altercentric ToM, with the hypothesis that one's own knowledge or ignorance serves as an important contextual element, determining whether and how one attributes mental states to those around them.

Experiments

Five experiments, in addition to an exploratory pilot study, were conducted to test the effects of egocentric knowledge or ignorance upon the execution of altercentric theory of mind.

Experimental Paradigm

In order to investigate the effects of egocentric knowledge and ignorance, an experimental paradigm was devised in which participants were asked to make predictions about the behavior of a character (see Figure 1). This task made use of a simple visual scene, in which three baskets were placed on top of a table, and a character stood behind the table facing the viewer.

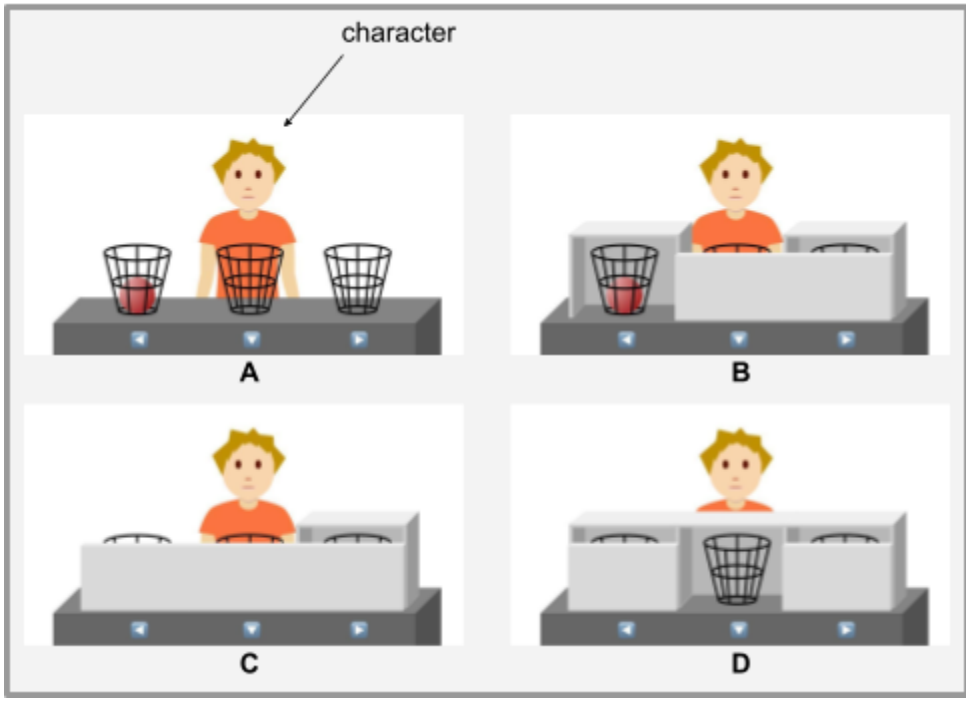

Figure 1. Example stimuli from Experiment 1. **A)** Both the participant and the character can infer the ball's location. **B)** The participant can infer the ball's location, but the character cannot. **C)** The participant cannot infer the location of the ball, but the character can. **D)** Neither the participant nor the character can infer the ball's location.

A single ball was hidden within the scene, and the participant was asked to predict where the character would search for the ball. In some trials, the location of the ball could be seen (or inferred) by both the character and the experimental participant (Figure 1, A). In other trials, the ball's location was known by the participant, but not the character (Figure 1, B). And in some trials, the ball's location was known to the character, but not to the participant (Figure 1, C), or known to neither (Figure 1, D). Thus, in this paradigm, experimental subjects were asked to make ToM-dependent judgements under cases of egocentric knowledge or egocentric ignorance, about a character who was reasoning from a position of altercentric knowledge or altercentric ignorance. This task thereby allowed for an investigation of the ToM-dependent behaviors of

experimental subjects when reasoning from positions of egocentric knowledge or egocentric ignorance.

General Hypothesis

This project is motivated by the general hypothesis that egocentric knowledge or ignorance states are vital indicators of whether, and how, one engages in altercentric ToM. In particular, one's knowledge or ignorance will have impacts upon their interest in the perspectives of others, and their expectations for the knowledge or ignorance of others. In the experimental paradigm used in this project, participants' predictions about where the character will look for the ball will be different when the participants do, or do not, know the ball's actual location. This difference may reflect adaptive preferences, which allow us to efficiently allocate our cognitive resources to make the most of our ToM with the goal of learning things from those in our environments. This hypothesis is the primary focus of experiments 1-4 of this project.

Exploratory Pilot Study

An exploratory pilot study was designed to test the present experimental paradigm, ensuring that it would be properly interpreted by participants, and also to allow for initial exploratory analyses. Exploratory analyses were planned to investigate the effects of egocentric knowledge, or ignorance, upon the predicted behavior of the character, within the task. In particular, these analyses investigated whether, when reasoning from positions of ignorance, participants were especially interested in positions to which the character had visual access. If so, participants may preferentially expect the character to search in these locations.

Methods

Stimuli and Procedures The experiment was administered as an online survey, created using the JsPsych JavaScript library. Participants were familiarized with the scene (shown in Figure 5, Pilot Stimuli) and introduced to the character as 'Sam.' They were told that their task, throughout the experiment, would be to predict where Sam would search for a red ball as it was hidden in one of the three baskets, and that they would receive a monetary bonus for each correct response. Participants then completed 10 practice trials, predicting Sam's behavior, before completing a 2-trial comprehension check. If the comprehension check was failed, participants repeated the instruction and practice components of the experiment before repeating the test.

The main body of the experiment consisted of 144 trials. In each trial, participants were presented with a stimulus showing the character, Sam, the three baskets, and a configuration of occluders, and predicted which of the three baskets Sam would choose to search for the ball. They received feedback about whether their response was correct, or incorrect, with a red box indicating the actual location Sam picked. In cases in which the location of the red ball was visible to Sam, or in which Sam could infer the ball's location, this 'correct' location was simply the location of the ball. In cases in which Sam could not infer the ball's location, or the ball's location was unknown, the 'correct' location was randomly selected from each of the possible locations in which Sam might rationally expect the ball to be. At the conclusion of the experiment, participants completed a demographic questionnaire and received debriefing materials.

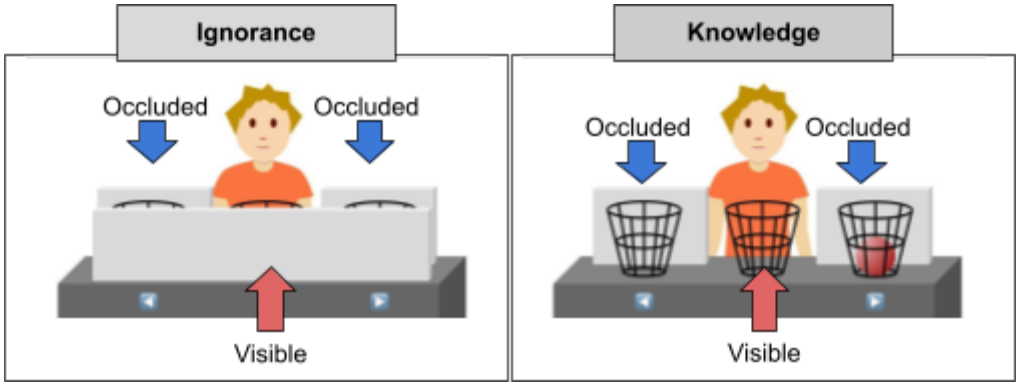

Figure 2. Example stimuli from the Pilot are shown. In the Participant Ignorant trials, the participant could not infer the location of the ball, while in Participant Knowledgeable trials, the participant could. This figure also highlights positions which are occluded, or visible, from the perspective of the character. In the Pilot, in Participant Ignorant trials, participants were more likely to select positions which were visible to the character.

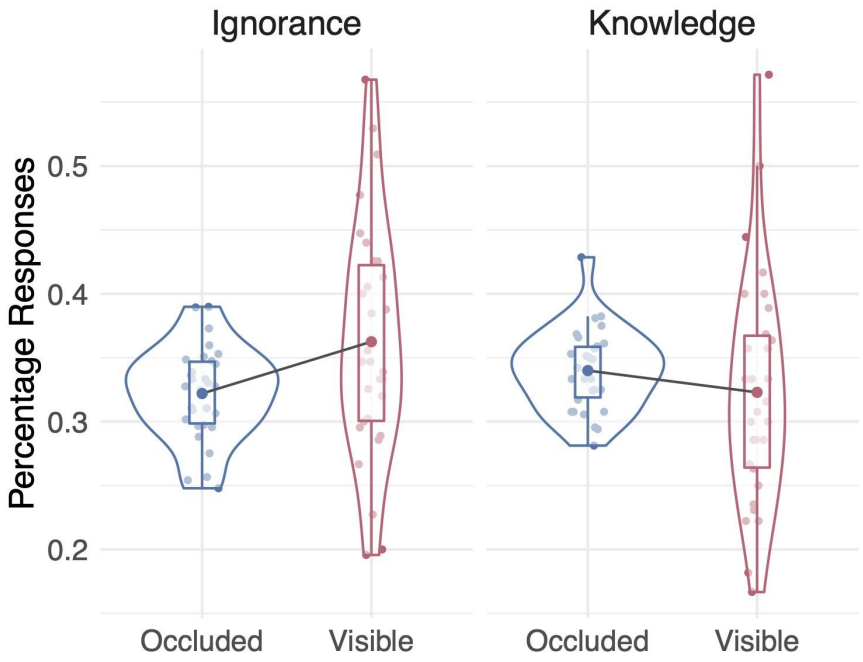

Figure 3. Participants' responses in the pilot study, in cases in which the participants could ('Knowledge') or could not ('Ignorance') infer the location of the ball. In cases of ignorance, participants were more likely to select positions which were visible to the character. In this experiment, there is a baseline difference in the number of positions which are visible, or occluded, from the character. This figure corrects for this baseline difference by displaying the percentage of occluded or visible positions which were selected, as compared to those that were available (i.e. the number of responses made, relative to the baseline availability of those responses).

Participants An arbitrary sample size of 30 participants was selected. Participants were recruited from Prolific, and were all adults, from the U.S.A., who were fluent in English and who had an approval rating of at least 97% with a minimum of 1000 prior submissions (age: M = 44, gender: 10 Female, 20 Male, 1 Non-binary). Participants were compensated at a base rate of $9/hr for this 10-minute experiment, and received a performance bonus of up to an additional $6/hr, with a target total compensation of $12/hr.

Results In cases in which there was a 'correct response,' that is, both the participant and Sam could infer the location of the ball, and there was, therefore, a single rational position in which Sam could be expected to search for the ball, participants had high accuracy in selecting this correct location (mean accuracy: 89%, sd: 11%). Participants had a tendency to select the position of the ball, even when the character had no way of inferring this location (Figure 4). Interestingly, as shown in Figure 3, participants' knowledge states had an influence upon their expectations for where Sam would search for the ball. When participants were reasoning from positions of ignorance (they could not infer the location of the ball) they expected Sam to search within locations which were visually available to Sam. This effect was validated analytically by a comparison of generalized linear models, using the lme4 library. A model predicting whether participants' selections were visible to Sam, using the participants' knowledge state and the number of positions visible to Sam as fixed effects, and the participants' ids as random effects:

Formula = (response_visible ~ participant_knowledge + number_visible_positions +
(participant_knowledge + number_visible_positions|participant_id))

was compared to another generalized linear model which was identical except for its omission of the fixed effect capturing participants' knowledge states, using the anova() function,

$\chi^2(1) = 5.24,\ p = 0.022.$ This result indicates that participants' knowledge states had a significant impact upon their preferences when predicting the character's behavior. In order to test the significance of this effect within cases of egocentric ignorance, an additional analysis was conducted on participant-level summaries, including only cases in which the participant was ignorant as to the location of the ball. A linear mixed-effects model was used to predict the frequency of participants' responses based upon whether those responses were altercentrically occluded or visible, using participants' ids as random effects:

Formula = (number_responses ~ response_visible + (1|participant_id))

This model was compared to a null model, which contained only the random effects. We found this effect to be highly significant, $\chi^2(1) = 111.82,\ p < 0.001$. This result indicates that, when reasoning from positions of ignorance, participants were significantly more likely to select positions which were within the visual access of the character, as shown in Figure 3.

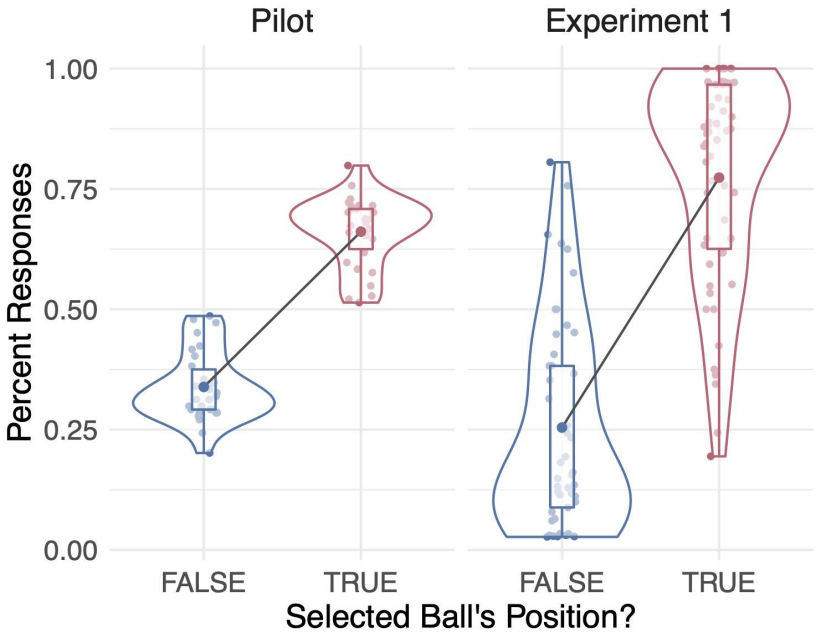

Figure 4. Participants' preferences to select the actual position of the ball, when predicting the character's behavior in the pilot study, and in Experiment 1, shown as the percentage of responses which did (TRUE) or did not (FALSE) select the ball's position. This figure shows only cases in which the participant could infer the location of the ball, but the character could not (so there was no single rational location in which the participant could be expected to search). As shown, participants were more likely to select the ball's position than any other position in these trials. This effect is particularly striking, as only one out of three positions contained the ball in each trial.

Discussion This exploratory pilot study uncovered an interesting effect—when using the present paradigm, participants' knowledge states influenced their predictions about where Sam would search for the ball. When participants were ignorant as to the location of the ball (Figure 2, Ignorance), they were more likely to select positions to which Sam had visual access. This result could be explained by an intuitive, and useful, tendency. Perhaps, when individuals are ignorant as to a salient fact in their environment, they become particularly interested in the privileged perspectives of those around them. Within the context of this experimental paradigm, this theory would lead to the prediction that, when participants are reasoning from positions of

ignorance (they don't know where the ball is), they become more interested in the locations which Sam can see, but they (the participant) cannot. This tendency would explain the observed effect, and would be advantageous for the purpose of learning information from the individuals in one's environment.

Of course, there are complications in interpreting the results of this preliminary pilot study. For instance, in this pilot study, there were baseline differences in the number of positions which were visible to the character, or not visible to the character. These baseline differences were accounted for within the statistical models used, and within the visualization of the data, but could still have behavioral effects not easily accounted for by these measures.

Additionally, as pointed out by colleagues, there was some ambiguity within the visual stimuli used in this pilot. It is possible that some participants interpreted the character as being able to see over the occluders used in the pilot (for an example of the pilot stimuli, see Figure 2). This potential issue was addressed in subsequent experiments.

Finally, the results of this pilot study would be strengthened by the replication of the observed effect, with similar but non-identical stimuli. Each of these issues is addressed by Experiment 1 of this project.

Experiment 1

Experiment 1 had the primary goal of replicating the results of the exploratory pilot study, with pre-registered predictions. Additionally, this experiment corrected the baseline differences in the frequency with which positions were visible, or not visible, to the character. Finally, this experiment used adapted stimuli, which corrected any ambiguity in the visual perspective of the

character. Occluders encased the basket on the top and sides to ensure that the character could not have visual access to occluded baskets. Predictions for this experiment, as well as additional information about the planned stimuli and procedures, are available at:

https://archive.org/details/osf-registrations-chskb-v1

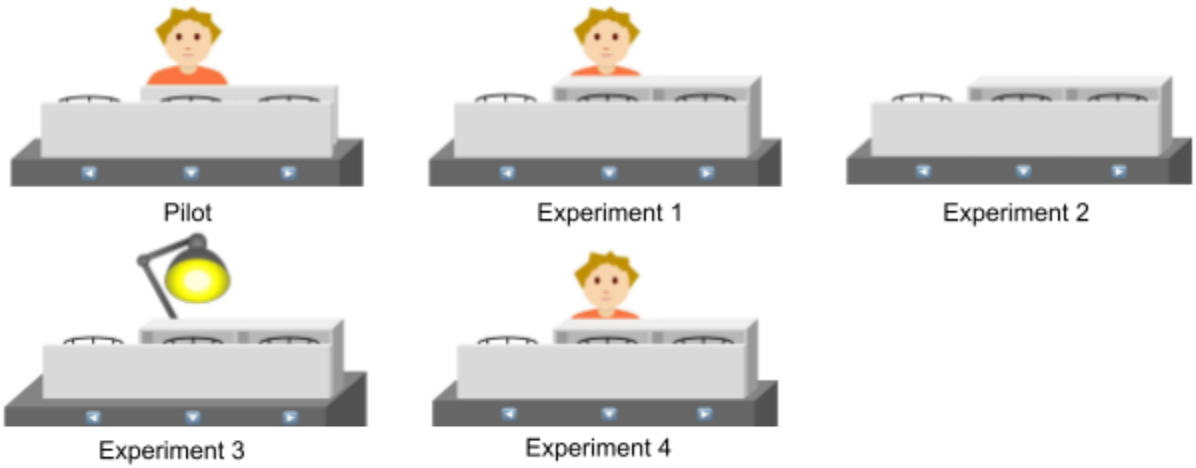

Figure 5. Sample stimuli from each experiment, and the exploratory pilot, are shown.

Methods

Stimuli and Procedures The stimuli and procedures for Experiment 1 closely resembled those of the Exploratory Pilot. The stimuli were adapted slightly from the pilot, to address concerns that the character, Sam, could be interpreted to have visual access to some of the baskets. In the stimuli for Experiment 1, the occluders eliminated this possibility, by enclosing the baskets on four sides. Again, participants received instructions, followed by a 10-trial practice phase, and a 2-trial comprehension check. The main body of the experiment consisted of 144 trials, with participants receiving feedback after each trial. In each trial, there was an equal

probability of the ball being in each of the three baskets, and a 50% probability that each basket would be occluded or not occluded from the perspective of the participant, and the perspective of the character. Participants again received a monetary bonus for each correct response made within the main body of the experiment. At the conclusion of the experiment, participants completed a demographic questionnaire, and received debriefing materials.

Participants This experiment had a target sample size of 55 participants. This sample size was based upon a post-hoc power analysis, conducted on the exploratory pilot study. In the pilot, an anova analyzing participant-level averages found a moderate effect size (f=0.27) when using egocentric ignorance as a predictor for the selection of positions occluded, or not occluded, from the perspective of the character. Our power analysis then suggested a sample size of 55 participants, for a target power of 80% at a significance threshold of p=0.05. Participants were recruited from Prolific, and were all adults from the U.S.A., who were fluent in English, had an approval rating of at least 97% with a minimum of 1000 prior submissions, and who had not participated in any prior experiments in this project (age: M = 42, gender: 30 Female, 25 Male). Participants were compensated at a base rate of $9/hr for this 10-minute experiment, and received a performance bonus of up to an additional $6/hr. The anticipated average rate of payment was $12/hr.

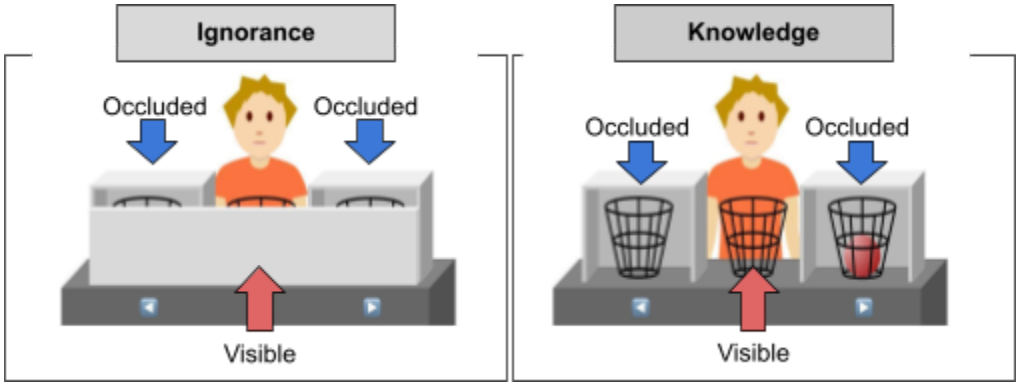

Figure 6. Example stimuli from Experiment 1 are shown. In the Participant Ignorant trials, the participant could not infer the location of the ball, while in Participant Knowledgeable trials, the participant could. This figure also highlights positions which are occluded, or visible, from the perspective of the character. In Experiment 1, in Participant Ignorant trials, participants were more likely to select positions which were visible to the character.

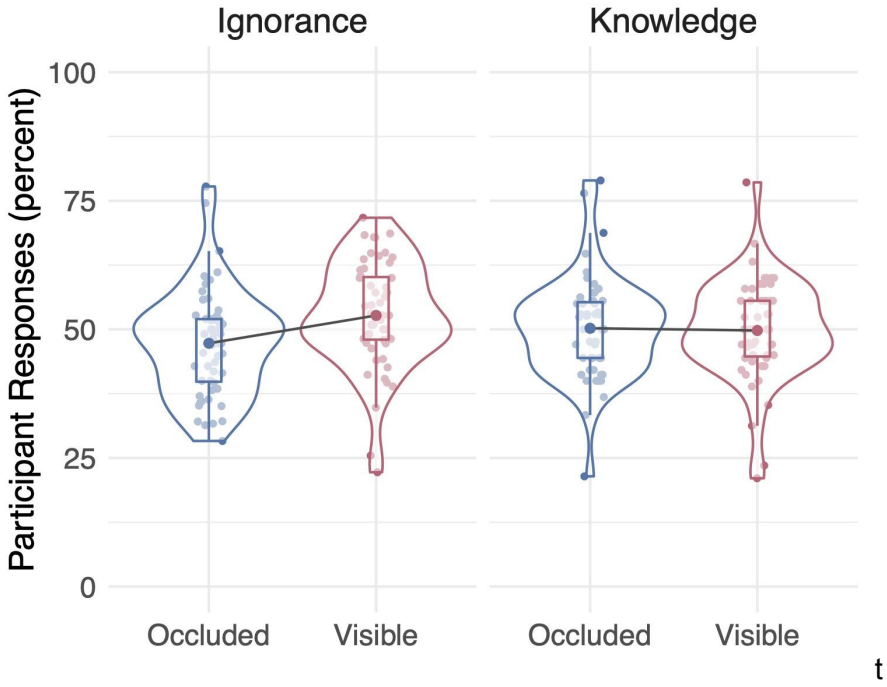

Figure 7. Participant responses in Experiment 1, in cases in which the participants could ('Knowledge') or could not ('Ignorance') infer the location of the ball. In cases of ignorance, participants were more likely to select positions which were visible to the character. Responses are shown as percentages *within* 'Ignorance' and 'Knowledge' conditions. See Figure 6 for examples of the cases illustrated in this plot.

Results In trials where both the participant and the character could infer the location of the ball, and there was therefore a rational position which the character should have selected, participants again had high accuracy in selecting the correct position (mean accuracy: 91%, sd: 13%). This result was consistent with the pilot study. Participants also, once again, had a preference to select the position of the ball, even when the character could not infer this location (Figure 4). As predicted, in Experiment 1, participants' knowledge states again had an influence upon their altercentric predictions. This effect was examined analytically using the same strategy employed in the Exploratory Pilot analyses. Comparison of a generalized linear model including the fixed effects of participant knowledge state, and number of altercentrically occluded positions, to a model containing only the number of altercentrically occluded positions, showed that this effect was significant, $\chi^2(1) = 5.33,\ p = 0.021$. Figure 7 shows this effect. Participants were again more likely to select positions which were visible to the character when they were reasoning from positions of ignorance, as compared to positions of knowledge, and this effect was once again shown to be significant, using a linear mixed-effects model where the altercentric access of responses (i.e. occluded or visible) was used to predict the participant-level frequency of those occluded or visible responses $\chi^2(1) = 7.62,\ p = 0.006$. These findings replicate those of the exploratory pilot study.

Discussion The findings of Experiment 1 replicate those of the Exploratory Pilot study, within a paradigm with no baseline differences in the number of positions which are occluded, or visible, from the character's perspective. Additionally, the improved disambiguated stimuli remove the possibility that some participants interpreted the character as being able to see the

baskets over the occluders. This concern is further addressed by the high accuracy in both the pilot and Experiment 1, in trials in which there was a rational position for the participants to select. This high accuracy indicates that the intended manipulations of the character's apparent perspective were successful.

Again, the results of this experiment indicated that participants' egocentric knowledge states had a significant influence upon the ways in which they participated in altercentric theory of mind judgements. Specifically, while there was no positive or negative preference to select altercentrically visible positions when participants were reasoning from positions of knowledge, when participants were reasoning from positions of ignorance, they showed a preference for selecting altercentrically visible positions. This preference may be evidence that participants are especially interested in the privileged knowledge states of those around them, in cases where other people could have valuable knowledge that they, the participant, do not.

But of course, there are also compelling competing explanations for this observed result. Participants' egocentric knowledge states might be expected to produce differing participant responses, even without the presently suggested adaptive preference. For example, participants may have a baseline preference to select positions to which the character has visual access. In cases in which participants know the actual location of the ball, this baseline preference may be overridden by the much stronger preference to select the actual position of the ball (a good strategy when trying to quickly choose the location where Sam could plausibly search for the ball). If this were the case, the observed effect may actually reflect 1) a preference by participants to select the location of the ball and 2) a slightly weaker preference to select positions which are visible to the character, owing to the 'low-level' features of the scene (i.e. visual or spatial features not relevant to ToM). For that matter, there might be other low-level visual features of

the present stimuli (for instance, the baseline attentional differences between positive and negative space), which could elicit this effect. Experiment 2 is designed to address these low-level confounds.

Experiment 2

As discussed, the procedures of Experiment 1 did not control for the effects of the differing low-level visual features of stimuli, in different conditions. Experiment 2 serves as a control study, validating the results of Experiment 1 by testing the effects of the low-level visual features of the present stimuli. It also uncovers participants' baseline assumptions about the actual location of the ball, when this location is not known.

Methods

Stimuli and Procedures The stimuli for Experiment 2 were identical to those used in Experiment 1, except that no character was depicted within the scene (see Figure 5). Participants were familiarized with the scene, and told that their task, throughout the experiment, would be to select the basket which contained the red ball, and that they would receive a monetary bonus for each correct response. Again, after receiving the instructions, participants completed a 10-trial practice phase, before completing a 2-trial comprehension check.

The main body of the experiment consisted of 144 trials. In each trial, participants were presented with a stimulus depicting the three baskets with the ball in one basket and a configuration of occluders. Participants selected the basket which contained, or was most likely to contain, the red ball, and received feedback about the accuracy of their response. A red

rectangle showed which of the baskets contained the ball. In cases in which the ball was visible, or the location of the ball could be inferred, this 'correct response' was simply the ball's location. When the location of the ball could not be inferred, this location was randomly selected from each of the possible locations. At the conclusion of the experiment, participants completed a demographic questionnaire, and received debriefing materials.

Participants Consistent with Experiment 1, this experiment had a target sample size of 55 participants. Participants were recruited from Prolific, and were all adults from the U.S.A., who were fluent in English, had an approval rating of at least 97% with a minimum of 1000 prior submissions, and had not participated in any prior experiments in this project (age: M = 42, gender: 28 Female, 26 Male, 1 Non-binary). Participants were compensated at a base rate of $9/hr for this 10-minute experiment, and received a performance bonus of up to an additional $6/hr. The anticipated average rate of payment was $12/hr.

Results In cases in which participants could infer the location of the ball, they had very high accuracy in selecting that location (mean accuracy: 97%, sd: 1.9%). Again, in Experiment 2, participants' egocentric knowledge states had a significant impact upon their responses $\chi^2(1) = 8.37,\ p = 0.004$. This effect is shown in Figure 8. However, this effect in Experiment 2 was in the opposite direction of Experiment 1. In Experiment 2, when participants were reasoning from a position of ignorance, they were more likely to select positions which were occluded from behind the table (which would be the character's perspective in Experiment 1). Using the same analysis of a linear mixed-effects model used in the pilot and Experiment 1, this effect was shown to be significant $\chi^2(1) = 32.95,\ p < 0.001$.

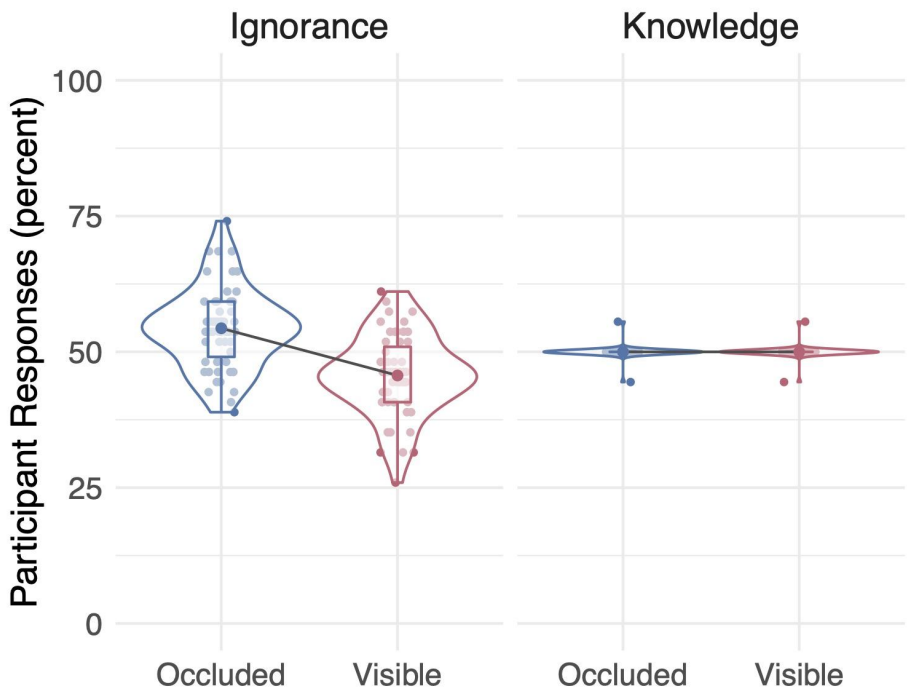

Figure 8. Participant responses in Experiment 2, in cases in which the participants could ('Knowledge') or could not ('Ignorance') infer the location of the ball. In cases of ignorance, participants were more likely to select positions which would have been occluded to the character, if the character were present. Percentages of responses which selected occluded or visible positions are shown, within cases of participant Ignorance and Knowledge.

Discussion The results of Experiment 2 validate those of Experiment 1, by testing participants' 'baseline' preferences in the absence of a human character within the scene. These results show that, in the absence of a character, participants' preferences are opposite those of Experiment 1. That is, in Experiment 2, when ignorant as to the ball's location participants were more likely to select positions which had occluders behind them than positions which did not. The results of Experiment 2 suggest that the effect observed in Experiment 1 cannot be explained by low-level features within the scene, owing to the differing placement of occluders between

conditions. And it cannot be explained by a general expectation that the ball will be hidden in more visually accessible positions, because as found in Experiment 2, when simply asked to guess the location of the ball, participants show the opposite pattern. So, low-level effects influencing participants' expectations concerning the ball's location, or influencing the salience of various aspects of the scene, would seem to contribute to a behavioral pattern *opposite* that observed in Experiment 1.

However, a slightly more sophisticated version of this concern, which relies upon different 'low-level' influences such as attentional cues, is still available to the objector. It has been suggested that the directional features of human faces may direct attention in ways that sometimes resemble, but ultimately do not constitute, theory of mind (Heyes, 2014; Santiesteban, 2014). While these powerful attentional effects may facilitate the fluent execution of theory of mind (for instance, helping us to notice the things that matter most to ToM), these processes do not, themselves, implicate ToM. It remains possible that, in Experiment 1, participants responded to the directional features of the character. This response may have caused a preference to attend to the parts of the scene to which the directional cue (the character) had unobstructed access. That is, perhaps the character's face 'pointed to' the areas in the scene which were not occluded from the character's view, but this attentional effect had nothing to do with ToM. Experiment 3 investigates this alternative explanation.

Experiment 3

Experiment 3 controls for the directional features of the character within Experiment 1. This is accomplished through the use of a lamp, which casts light in a pattern resembling the

visual field of the character. As such, the lamp emulates the directional features of the character, without being a viable candidate for the attribution of mental states (Schurz, 2015). If the results of Experiment 1 were owed to the directional features of the character, it is expected that Experiment 3 would replicate these results. These predictions, along with additional information about the planned stimuli and procedures, are pre-registered at:

https://archive.org/details/osf-registrations-uyvg3-v1

Methods

Stimuli and Procedures The stimuli for Experiment 3 were identical to those used in Experiment 1, except that the character, Sam, was replaced with a lamp, with the position of the head of the lamp in Experiment 3 resembling the position of Sam's head in Experiment 1. The field of light cast by the lamp was intended to resemble the visual field of the character, Sam, from Experiment 1.

The procedure for Experiment 3 was identical to that of Experiment 2. After receiving instructions, and an attention check, participants completed 144 trials in which they selected the basket which was most likely to contain the ball. They received feedback, after each trial, showing the ball's correct location, and received a monetary bonus for each correct response. At the conclusion of the experiment, participants completed a demographic questionnaire and received debriefing materials.

Participants Consistent with prior experiments, this experiment had a target sample size of 55 participants. Data from one participant was not recorded by the data collection system, resulting in a final sample size of 54. Participants were recruited from Prolific, and were all adults from the U.S.A., who were fluent in English, had an approval rating of at least 97% with a

minimum of 1000 prior submissions, and had not participated in any prior experiments in this project (age: M = 40, gender: 29 Female, 24 Male, 1 Not Specified). Participants were compensated at a base rate of $9/hr for this 8-minute experiment, and received a performance bonus of up to an additional $6/hr. The anticipated average rate of payment was $12/hr.

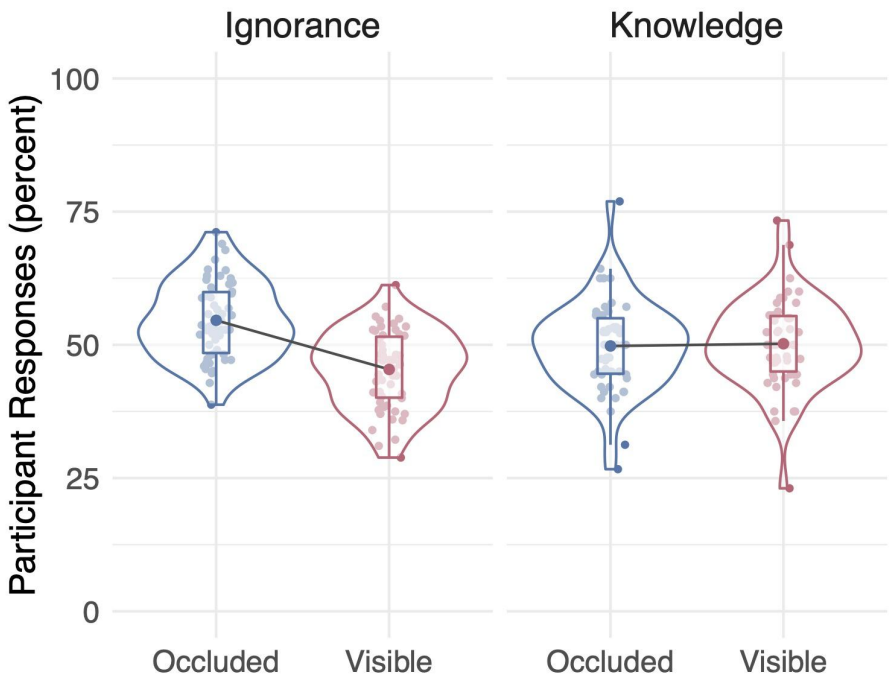

Figure 9. Participant responses in Experiment 3, in cases in which the participants could ('Knowledge') or could not ('Ignorance') infer the location of the ball. In cases of ignorance, participants were more likely to select positions which were not within the illumination field of the lamp (would have been occluded to the character). Percentages of responses which selected occluded or visible positions are shown, within cases of participant Ignorance and Knowledge.

Results Participants once again had high accuracy in cases in which it was possible for them to infer the location of the ball (mean accuracy: 97%, sd: 1.1%). Consistent with Experiment 2, and as shown in Figure 9, when participants were reasoning from a position of

ignorance, they were more likely to select positions which were occluded from behind the table (which would be the character's perspective in Experiment 1) $\chi^2(1) = 32.63,\ p < 0.001$. The effect of participants' knowledge states upon their selection of visible or occluded positions (from the position of the lamp) was significant $\chi^2(1) = 7.58,\ p = 0.006$. So, in the presence of a nonhuman directional stimulus, the effect of egocentric knowledge was similar to that found in the absence of a directional stimulus in Experiment 2, and was opposite in direction to that found in Experiment 1.

Discussion Experiment 3 replicated the results of Experiment 2, using a nonhuman directional stimulus to emulate the directional effects of the human character depicted in Experiment 1. Once again, the results of Experiment 3 suggest that low-level attentional effects cannot explain the results of Experiment 1. That is, the preference to select positions which are visible to the human character in Experiment 1, when reasoning from positions of egocentric ignorance, were not reproduced using an inanimate stimulus that matched the directional features of the human character. Instead, these results again suggested that subjects' baseline preferences, in the absence of a target for theory of mind, are opposite those which are observed when participants are asked to perform ToM. When participants do not know the location of a ball, and are reasoning about where another individual will look for that ball, their prediction concerning the other's behavior is different from their general prediction about the actual location of the ball. These results further support the notion that the results of Experiment 1 depend upon participants' specific expectations when reasoning about the mental states of others from positions of uncertainty. When reasoning from positions of ignorance, participants become

especially interested in the perspectives of others when those perspectives might help participants to ‘fill in the gaps’ in their own knowledge.

Experiment 4

Experiments 2 and 3 show that the low-level features of the present stimuli are not sufficient to account for the results of Experiment 1. Even still, one might be concerned that the visual stimuli used in Experiments 2 and 3 are non-identical to those which were used in Experiment 1. So, even if they are similar in relevant respects to the stimuli used in Experiment 1, they may not account for the directional or spacial features of those stimuli. To address this concern, Experiment 4 of this project used identical stimuli to Experiment 1, including a human character. However, once again, participants were asked to respond according to their own expectations concerning the location of the ball, and not those of another individual.

This experiment was expected to replicate the findings of Experiments 2 and 3, despite the use of stimuli identical to those of Experiment 1. This finding would address concerns about the low-level features of the stimuli used in Experiment 1. Additionally, this finding would inform the level of ‘automaticity’ implicated in the results of Experiment 1. Within the theory of mind literature, a distinction has been made between ‘automatic’ processes, which function mandatorily irrespective of factors such as participants’ goals or intentions within a task, and processes which are merely ‘spontaneous,’ which happen efficiently and effortlessly, but may not happen involuntarily (O’Grady, 2020). We hypothesized that the key effect from Experiment 1 was not automatic, and was dependent upon participants’ intentional consideration of the character’s perspective, and not merely the presence of the character. The predicted results of

Experiment 4, as well as additional information about the stimuli and procedures are pre-registered at: https://archive.org/details/osf-registrations-khc92-v1

Methods

Stimuli and Procedures The stimuli used in Experiment 4 were identical to those used in Experiment 1. They depicted the character, standing behind a table, with three baskets on top of the table. Just as in Experiment 1, in each trial, a ball was hidden in one of the three baskets, and occluders partially obscured the baskets from both Sam's perspective, as well as from the participants' perspective. However, the procedure of Experiment 4 was identical to those of Experiments 2 and 3. That is, after receiving instructions, and an attention check, participants completed 144 trials in which they selected the basket which was most likely to contain the ball. They received feedback, after each trial, showing the ball's correct location, and received a monetary bonus for each correct response. At the conclusion of the experiment, participants completed a demographic questionnaire and received debriefing materials.

Participants Consistent with prior experiments, this experiment had a target sample size of 55 participants. Participants were recruited from Prolific, and were all adults from the U.S.A., who were fluent in English, had an approval rating of at least 97% with a minimum of 1000 prior submissions, and had not participated in any prior experiments in this project (age: M = 42, gender: 29 Female, 24 Male, 1 Non-binary, 1 Not Specified). Participants were compensated at a base rate of $9/hr for this 8-minute experiment, and received a performance bonus of up to an additional $6/hr. The anticipated average rate of payment was $12/hr.

Results Once again, participants had high accuracy in correctly selecting the ball's location when it was possible to infer this location (mean accuracy: 97%, sd: 2.6%). Consistent

with Experiments 2 and 3, when participants were reasoning from a position of ignorance, they were more likely to select positions which were occluded from the perspective of the character $\chi^2(1) = 22.72,\ p < 0.001$. This result is shown in Figure 10. The effect of participants' knowledge states upon their selection preferences was significant $\chi^2(1) = 5.32,\ p = 0.021$. These results replicate those of Experiments 2 and 3, while using visual stimuli which are identical to those from Experiment 1.

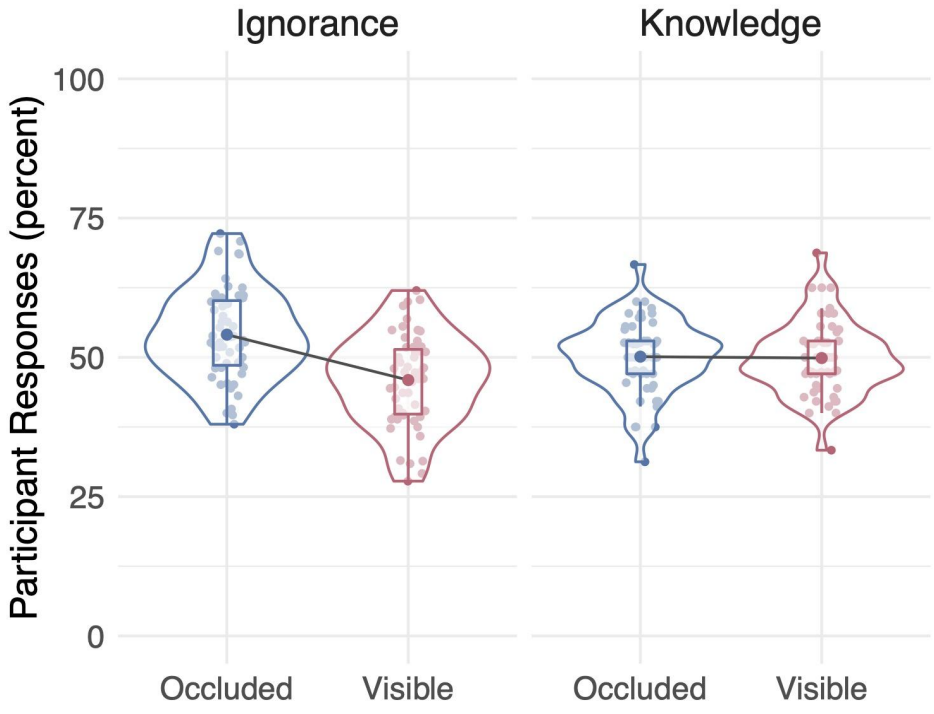

Figure 10. Participant responses in Experiment 4, in cases in which the participants could ('Knowledge') or could not ('Ignorance') infer the location of the ball. In cases of ignorance, participants were more likely to select positions which were occluded from view for the human character. Percentages of responses which selected occluded or visible positions are shown, within cases of participant Ignorance and Knowledge.

Discussion In Experiment 4, participants were presented with stimuli identical to those from Experiment 1, but were given different instructions. Their responses were consistent with those from Experiments 2 and 3, showing a pattern of results opposite from Experiment 1. So, even in the presence of a human character, when reporting on their own expectations and reasoning from positions of ignorance, participants expect the ball to be located in positions which are hidden from the character's view. This pattern is opposite that found in Experiment 1. This result persuasively dissociated the results of Experiment 1 from the low-level features of the visual stimuli used, and participants' general expectations about the position of the ball within this paradigm. The key effect from Experiment 1 must, it seems, be owed to participants' specific activity of reasoning about the character's mental states.

This result also informs the level of 'automaticity' within the cognitive processes underlying the key effect in Experiment 1. As found elsewhere in the ToM literature, it seems that this response is not an 'automatic' response to the stimuli (O'Grady et al, 2020). While the present work is not designed to further determine the levels of explicit cognitive control, demand, or awareness underlying this effect, this issue provides an interesting route for future research.

General Discussion

The results of the pilot study and Experiment 1 reveal a replicable effect; when reasoning from positions of ignorance, participants expect the character to search for the ball in locations to which he has visual access. And, as shown in Experiments 2-4, this effect cannot be explained by the attention-directing features of the stimuli. Further, when asked in separate experiments, participants do not themselves expect the ball to be located in altercentrically visible locations. So then why do participants expect the character to search for the ball in these locations? I will briefly discuss three explanations for this result: 1) that participants do not believe that the ball is in these locations, but nonetheless expect participants to search there, 2) that these responses are guided by case-specific heuristics, and 3) that these responses result from an allocation of attentional resources, which facilitates ToM. I will argue that the third and final of these explanations best accounts for the present observations.

Explaining the Present Effect

One potential explanation for this result is that, all the while, participants have consistent expectations as to the location of the ball, but they predict that the character will search for the ball in the incorrect location in the key trials of Experiment 1. The observed results can then be explained by 1) the general expectation, when reasoning from positions of ignorance, that the ball will be in more visually accessible locations, and 2) the expectation that the character will select incorrect locations when the character is ignorant. But this explanation is incompatible with the observation that, in general, participants expect the character to make correct predictions even when the character has no principled way of knowing where the ball is (Figure 4). It would

be very strange if participants expected the character to guess incorrectly, only in the critical trials of Experiment 1.

Alternatively, it could be the case that participants' expectations concerning the location of the ball genuinely differ between trials. Participants may, for example, operate under heuristics such as "when I don't know where the ball is, it must be hidden in a maximally inaccessible position" and "when I am ignorant as to a salient particular, and someone else has pertinent knowledge, that pertinent knowledge must be informative as to the salient particular." Other similar heuristics could be proposed, which would be explanatorily equivalent given the very sparse present data. Uncovering such assumptions would surely be interesting. But the problem with explanations along this line is that they make quite a lot of quite a little. It may be the case that participants do have heuristics underlying their expectations of the ball's location, and that these heuristics produce opposite patterns of results in Experiment 1 as compared to Experiments 2-4. But there are simpler, and more unifying accounts that can explain the present observations. Empirical means of dissociating the presence of these 'heuristics' from simpler alternative accounts will be discussed in *future directions.*

The simpler explanation of the observed findings is that the key effect in Experiment 1 depends upon the differing visual (attentional) salience of altercentrically occluded as compared to altercentrically visible locations within the scene. But, as shown in Experiments 2-4, this attentional effect is not a simple response to the visual stimuli used. Instead, it is a result of participants' goal-induced consideration of the character's visual perspective, and potential behaviors, as they reason from positions of ignorance. In other words, as participants explicitly consider the character's visual perspective with the goal of predicting where he will search for the ball, their attention is directed to the regions in their environment to which the character has

visual access. But this effect is not uniformly observed across all trials. Instead, this effect is pronounced only when participants do not, themselves, know the location of the object for which the character is searching.

Consider a slightly different high-level characterization of this explanation: When participants do know the location of the ball, and are asked to make predictions about where the character will search for the ball, they attend to the ball's location, because the character's intentions are aimed at this location, even if they don't have any rational means of ascertaining where in the scene that location is. On the other hand, when participants themselves do not know the location of the ball, and have the goal of predicting the behavior of the character, participants turn their attention to the parts of the scene which the character would act upon most readily, given a certain of several possible positions for the ball. The ball *might* be in these locations. Of course, it also might not. But if it is, the character will have privileged knowledge about the ball's location, and will surely select these positions. If it is not, the character will be in a similar position of ignorance as the participant, and will have to guess where the ball is (just as the participant would).

This explanation is, of course, highly speculative. This project constitutes a very cursory exploration of egocentric knowledge states upon the execution of altercentric ToM. But the latter proposed explanation describes a rational response to 'knowledge advantages' within this paradigm—when the participant knows something the character doesn't, or vice versa. In a task where the salient piece of information is the position of a ball, and the participant knows that precise position, it is clear to them that any additional knowledge held by the character is inconsequential. Why should the character's privileged knowledge matter if the participant knows it to be unhelpful? But when the participant doesn't know the location of the ball, and the

character *might,* the character's privileged knowledge is very important. In these cases, the participant should be attentive to the character's special knowledge.

It is easy to see how this contextually dependent attentional direction would be advantageous in cooperative or competitive environments. It isn't possible to attend to the mental states of everyone in one's surroundings. This kind of rampant 'automatic' ToM would be unnecessary and costly, and is not supported by prior research (O'Grady et al, 2020). Instead, our attention should be directed towards the perspectives of those who may have pertinent information, which can help to 'fill in the gaps' in our own knowledge. In this way, our own egocentric knowledge or ignorance can serve as a template for which information in our environment, and which perspectives among those in our surroundings, are worth considering.

Theory of Mind and Attentional Effects

It is worth recalling that Experiments 2-4 were motivated by concerns about attentional effects from the 'low-level' (i.e. spatial or directional, as opposed to agentive) visual features of the stimuli. These experiments found that such visual features could not explain the findings from Experiment 1. Evidence of 'attentional effects' within ToM research has often been viewed as debunking confounds, providing alternative explanations for empirical findings that do not rely upon 'genuine ToM' (Santiesteban et al, 2014). In some cases, it has been proposed that these attentional effects aid in the efficient execution of ToM, for example directing our attention in the direction of another's gaze, but that these attentional processes do not themselves constitute ToM (Heyes, 2014). The herein proposed explanation of Experiment 1 draws upon another sort of attentional direction, which is sensitive to the knowledge states of the participants,

and is guided by the perspective of the character. One might have good reason to wonder whether such a process should 'count' as an execution of ToM. Does this project belong in the ToM literature at all, or would it be better suited to the literature on human visual attention?

I've proposed that in this experiment, participants are responsive to their own egocentric knowledge or ignorance. But one can respond to their own knowledge states without directly forming 'representations' of these states. For example, when searching for the ball, one might know that the ball must be hidden in one of two locations, and know that it cannot be hidden in a third. But action on this information does not necessitate explicit representations of one's own knowledge states. We can respond to our knowledge states directly without 'metarepresentation' (i.e. representation of our knowledge states, which are themselves representations of states of affairs). So, our egocentric knowledge states might direct our attention, and influence whether and how we consider the mental states of others, without necessitating the sort of 'metarepresentation' characteristic of 'genuine ToM.' Likewise, the attentional effects derived from the perspective of the character within this paradigm are also not necessarily constitutive of ToM. That is, the direction of one's attention towards, or away from, the perspective of another person does not itself constitute an attribution of mental states to that person, or a representation of the content of their perspective. In short, the present findings needn't result from a process that should 'count' as ToM, according to the ordinarily applied standards (Bennett, 1978; Dennett, 1978; Harman; 1978). But I will argue that discussion of these standards—of what 'counts' and what doesn't—is not of primary importance for this project.

A careful re-articulation of the goals of this project should make progress towards answering this concern. Considerable prior ToM research has focused on edge cases—children, intelligent animals, large language models—with the goal of uncovering *whether or not* these

agents or systems are capable of ToM (Harman, 1978; Liszkowski et al, 2006; Kosinski, 2023). In these cases, it is worthwhile to develop a strict and careful description of 'what counts' as ToM, and to vigilantly pursue plausible explanations of behaviors that appear to demonstrate ToM, but may fall short of the specified standard. This project, on the other hand, deals with a case in which healthy adults are explicitly asked to predict another's behavior, and must do so in order to competently complete the task (which they indeed do successfully). It is obvious that theory of mind is demonstrated in this task. The aim is to characterize processes that facilitate the execution of ToM, within a paradigm where it is surely happening. The question is not *whether or not* but rather *how,* and in answering this question, it is clear that attentional effects which are informed by egocentric knowledge states, and increase the efficiency of altercentric ToM, are informative and relevant. It is important to this project that these attentional effects are not 'low-level'—that they do not capitalize on attentional phenomena which have nothing to do with ToM—and Experiments 2-4 attempt to address this concern. But the direction of attention, when uniquely responsive to ToM or perhaps necessary to its fluent execution, is precisely the kind of effect that is of interest.

This project thus urges consideration of a 'ToM system' just as visual processes rely upon a 'visual system.' An understanding of this system will necessitate thorough investigation at many levels. And while there will sometimes be cause for discussion of which processes in this system definitively demonstrate ToM, a narrow focus on this question, especially when investigating the nature of ToM as it happens in healthy human adults, may not always be fruitful. In other words, just as many investigations of psychological processes, such as vision, do not primarily focus on whether particular processes 'count as vision,' this project focuses on *how*

the ToM system operates and is not primarily concerned with drawing boundaries around which human behaviors should be included within this system.

Future Directions

A clear and prevailing explanation of the presently observed effect will necessitate further research. A promising direction for this work would probe participants' attentional preferences directly, for example, using eye-tracking as a measure for the salience of various parts of a visual environment, as participants consider the perspective of a character. If participants selectively attended to aspects of the scene within the character's visual field *only* when reasoning from positions of ignorance as to a relevant fact, this result would support the proposed attentional explanation of the present findings. This kind of direct investigation of attention could lend credibility to the more general attentional account, over one appealing to more specific heuristics.

Future research may also seek to uncover additional facts about the ways in which egocentric knowledge states influence the execution of altercentric theory of mind. For example, there may be different constraints upon the attribution of knowledge when one does, or does not, actually have access to that knowledge (Westra and Nagel, 2021). We might still understand that 'the character knows where the ball is' even if we, ourselves, do not. And these knowledge attributions may differ in kind from those made when we do know the position of the ball. These kinds of knowledge attributions, made from differing epistemic positions, provide an interesting and largely unstudied avenue for future investigations of the effects of egocentric knowledge states.

More broadly, future ToM research should consider the ways in which the 'theory of mind system' is reciprocally dependent upon our attentional and perceptual systems. Our epistemic states may influence the attentional mechanisms which allow us to fluently navigate social environments, just as our goal-induced consideration of others' perspectives allows us to allocate our attention effectively, and to update our own understandings of our environments advantageously. As is often the case, a clear-headed description of real-world ToM will need to pay tribute to these rich interdependencies, if it is to explain the efficiency and flexibility that makes human ToM so special.

Concluding Remarks

An interesting effect is uncovered in this project; when reasoning from positions of ignorance, participants appear to be especially interested in the visual perspective of a human character. This finding supports the idea that participants' underlying knowledge states influence the ways in which they consider the knowledge, or ignorance, of others. Future work will be required to uncover the specific mechanisms driving this effect. This further research will expand our understanding of human ToM, by investigating the ways in which the targeted direction of attentional resources may allow for fast and effective consideration of the minds of those in our environment.

References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”?. *Cognition*, *21*(1), 37-46.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences*, 1(4), 557–560.

Doherty, M. (2008). Theory of Mind: How Children Understand Others' Thoughts and Feelings. United Kingdom: Taylor & Francis.

Call, J., & Santos, L. R. (2012). Understanding other minds. In Mitani, J., Kappeler, P., Palombit, R., Call, J. & Silk, J. (Eds.), The evolution of primate societies (pp. 664–681). University of Chicago Press.

Carrington, S. J., & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human brain mapping*, *30*(8), 2313-2335.

Dennett, D. C. (1978). Beliefs about beliefs [p&w, sr&b]. *Behavioral and Brain sciences*, *1*(4), 568-570.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 459-473.

Drayton, L. A., & Santos, L. R. (2016). A decade of theory of mind research on Cayo Santiago: Insights into rhesus macaque social cognition. American Journal of Primatology, 78(1), 106–116.

Harman, G. (1978). Studying the chimpanzee's theory of mind. *Behavioral and Brain Sciences*, *1*(4), 576-577.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131-143.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibly to others' beliefs in human infants and adults. Science, 330, 1830–1834.

Liszkowski, U., Carpenter, M., Striano, T. & Tomasello, M. (2006). 12- and 18-month-olds point to provide information for others. *Journal of Cognition and Development*, 7(2), 173–187.

O'Grady, C., Scott-Phillips, T., Lavelle, S., & Smith, K. (2020). Perspective-taking is spontaneous but not automatic. *Quarterly Journal of Experimental Psychology*, *73*(10), 1605-1628.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? Science, 308, 255–258.

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and individual differences*, *43*(3), 541-551.

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., ... & Knobe, J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*, *44*, e140.

Phillips, J., & Norby, A. (2019). Factive theory of mind. Mind & Language, 36, 3–26.

Sally, D., & Hill, E. (2006). The development of interpersonal strategy: Autism, theory-of-mind, cooperation and fairness. *Journal of economic psychology*, *27*(1), 73-97.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people

see. *Journal of experimental psychology: human perception and performance*, *36*(5), 1255.

Santiesteban, I., Catmur, C., Hopkins, S., Bird, G., & Heyes, C. (2014). Avatars and arrows: Implicit mentalizing or domain-general processing?. Journal of Experimental Psychology: Human Perception and Performance, 40(3), 929.

Schurz, K. (2015). Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. NeuroImage (Orlando, Fla.), 117, 386–396.

Westra, E., & Nagel, J. (2021). Mindreading in conversation. *Cognition*, *210*, 104618.

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