Probing Nonhuman Primate Errors on False Belief Tasks to Explore the Evolutionary Roots of Theory of Mind

Amanda Royka*[1], Daniel Horschler*[1], Walker Bargmann[1], & Laurie Santos[1]

[1]Department of Psychology, Yale University *Denotes equal authorship credit

Abstract

Theory of Mind (ToM) is central to human social cognition, yet the roots of this capacity remain poorly understood. Both infants and nonhuman primates show inconsistent performance on false belief tasks, limiting our understanding of the representations that characterize the foundations of ToM. Here, we try to better understand this complex and often contradictory literature by dissecting these failures. Specifically, we focus on nonhuman primates' characteristic null performance on false belief tasks to test the circumstances under which they generate predictions about the behavior of an agent. Across three studies (n=419 subjects), we find that—despite succeeding on a closely matched control—rhesus monkeys fail to predict how agents with false beliefs will behave even when the agents perform highly unexpected, unlikely actions. We interpret this pattern of performance as evidence that monkeys may have no representation of another agent's past awareness once the scene changes outside of that agent's awareness, thus preventing them from making predictions about the agent's future actions. Overall, this work helps to move beyond the success/failure dichotomy typically used to assess ToM, and instead gives a more precise characterization of primates' signature limits in ToM, which we argue may also be shared with human infants.

Keywords: Theory of Mind; comparative cognition

Introduction

For decades, psychologists have sought to understand the roots of Theory of Mind (the capacity to understand and predict others' behaviors on the basis of inferred mental states; hereafter ToM; Premack & Woodruff, 1978; Wimmer & Perner, 1983), asking questions such as when are children able to represent others' mental states? And which ToM-like abilities are shared with our nonhuman primate relatives and which aspects are unique to humans? However, efforts to answer these questions have generated complicated and at times contradictory findings.

While young children and nonhuman primates (hereafter primates) consistently show the capacity to infer others' knowledge (sometimes argued to be a simpler state of awareness; see Martin & Santos, 2016) and make predictions about how those agents will subsequently behave (Arre et al., 2021; Bartsch & Wellman, 1995; Drayton & Santos, 2018; Flombaum & Santos, 2005; Hamlin et al., 2013; Hare et al., 2000, 2001; Holland & Phillips, 2020; Kaminski et al., 2008; Krachun et al., 2009; Luo & Johnson, 2009; MacLean & Hare, 2012; Marticorena et al., 2011; Martin & Santos, 2014; Melis et al., 2006; Santos et al., 2006; Wellman & Liu, 2004; Wellman & Woolley, 1990), studies have failed to find consistent evidence that young children and primates can represent and predict others' behaviors based on their false beliefs. For example, infants tested with nonverbal false belief tasks will, in some studies, pre-emptively look to where an agent falsely believes a desired object is located (as if correctly anticipating their actions; Garnham & Perner, 2001; Southgate et al., 2007; Surian & Geraci, 2012) and look longer when an agent does not act consistently with her false beliefs (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2009). Yet, in other tasks—including direct replications with increased sample sizes—infants fail to anticipate the false-belief-driven actions of others (Barone et al., 2022; Dörrenberg et al., 2018; Kulke et al., 2018) and do not look longer when agents with false beliefs search for an object in its true location (Dörrenberg et al., 2018; Powell et al., 2018). Similarly, primates fail to strategically take advantage of others' false beliefs in food competition tasks (Hare et al., 2001; Kaminski et al., 2008; Krachun et al., 2009) and do not look longer when an agent acts inconsistently with her false beliefs (Drayton & Santos, 2018; Horschler et al., 2019; Marticorena et al., 2011). However, primates sometimes pre-emptively look to where an agent falsely believes an object is located, suggesting that they may (under some circumstances) correctly anticipate actions motivated by false beliefs (Hayashi et al., 2020; Kano et al., 2019; Krupenye et al., 2016).

How can we move forward from this contradictory body of evidence? One approach is to focus on the specific errors that these populations make to better understand why primates and infants sometimes fail false belief tasks. Interestingly, there are two ways in which primates and infants *could* be failing these tasks. One possibility—which we refer to as the *multiple predictions account*—is that infants and primates have multiple predictions about what an agent with a false belief will do next (e.g., the agent may be equally likely to search in a location where they falsely believe an object to be or in a location where the object last was). Under this view, infants and primates *do* successfully track and represent some information about what the agent has seen in the past, but, in the end, think it is possible that the agent will search at either location. Alternatively, infants and primates might exhibit chance performance when making predictions about an agent with false beliefs because they have no expectations about how that agent will behave once her mental states are out of step with reality — a possibility we refer to as a *no prediction* account. Under this view, infants and primates have no expectations to be violated (and thus show null performance).

Distinguishing between these two explanations would help researchers to generate hypotheses about the representations that characterize ToM and to build better explanations for contradictory findings in the literature. Here, we take a first step towards understanding these failures by specifically focusing on primates.

Consider a typical primate violation of expectation false belief task (e.g., Drayton & Santos, 2018; Horschler et al., 2019; Marticorena et al., 2011) in which primates see a demonstrator witness a desirable reward being hidden inside one of two boxes. The demonstrator's view is then occluded as the reward moves back into its original box. Subjects then see the demonstrator reach either into the box where the reward is hidden or into the empty box. If primates expect someone with a false belief to search for an object where they believe it to be, then they should look longer when the agent reaches towards the object's true location, but primates do not show this pattern of looking; they instead look equally long at both outcomes.

Here, we adapt this typical design by adding a completely irrelevant hiding location. We reasoned that if primates generate *multiple predictions* about how an agent with a false belief will behave, then an agent searching in this irrelevant location should be surprising given that this behavior falls outside of the range of the predicted actions. In contrast, if primates have *no prediction* about how an agent with a false belief will behave, an agent who searches in a completely irrelevant hiding location should still be seen as unsurprising since primates have no expectation to be violated.

Study 1

Methods

Subjects We tested 148 rhesus macaque monkeys (*Macaca mulatta*) at the Cayo Santiago Field Station (60 females; mean age = 5.46 ± 3.05; see OSF Repository for complete demographic information). Other monkeys were approached but did not complete participation in the study because they were inattentive during critical parts of the study (n = 16), left the area (n = 16), were displaced by another monkey (n = 15), had previously participated in part of the study (n = 28), or due to an apparatus failure (n = 1). Decisions not to include a subject for this and all subsequent studies were made by the cameraperson who was blind to condition.

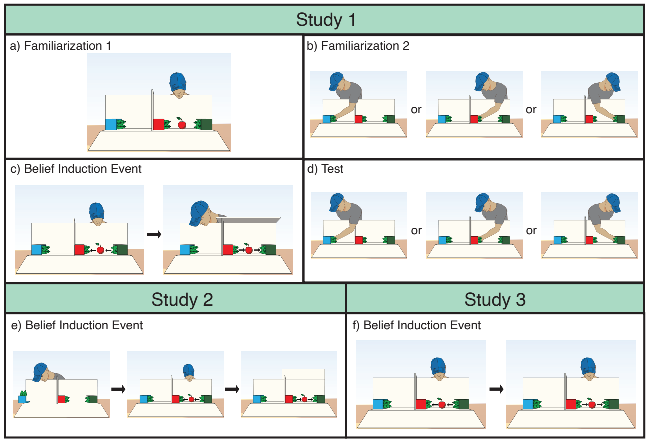

Figure 1: The procedure for all studies. a) In Study 1, subjects first saw the demonstrator staring at the apple in Familiarization 1. b) Then, in Familiarization 2, subjects saw one of three possible reaches depending on a blindly-assigned condition. c) The belief induction event next established the demonstrator's false belief about the location of the apple. d) In the test event, the demonstrator reached to the same box that they did in Familiarization 2, such that the demonstrator made one of three possible reaches: to the box that never held the apple (irrelevant reach), to the box where they last saw the apple (false-belief-consistent reach), or to the box where the apple actually was (reality-consistent reach). e) In Study 2, the demonstrator first observed that the irrelevant box was empty and then observed the apple move into one box, but did not see the apple move back to its original location. f) In Study 3, the demonstrator observed all of the apples movements and thus had a true belief about the apple's final location.

Apparatus We designed our false belief task to closely parallel ones used previously with this population (Horschler et al., 2019; Marticorena et al., 2011). Monkeys watched a series of trials in which an experimenter interacted with an apple on a small foamcore stage (36 in long, 5.5 in tall, and 6.5 in deep. The stage had a back (14 in high) and three small boxes (4 x 4 x 4 in) equally spaced along the stage. One of the boxes (hereafter the *irrelevant box*) was separated from the other two by a wall that was connected to the back of the apparatus. We used a small slit in the stage between the two other boxes (hereafter the *relevant boxes*) to slide an apple in between and into the two boxes surreptitiously. A front occluder (36 x 10.5 in) ran the length of the apparatus, allowing the experimenter to completely hide the stage from the monkeys' view. A smaller top occluder (21.5 x 8.5 in), was attached to the back of the apparatus which could block the experimenter's view (but not the subject's view) of the relevant boxes. The same apparatus was used in all subsequent studies.

Procedure Experimenters (demonstrator and cameraperson) opportunistically approached monkeys who were sitting calmly. The demonstrator knelt behind the apparatus 2 m away from the subject. A cameraperson stood behind the demonstrator to record the monkey's looking time. Each subject saw two familiarization trials and one test trial.

The *first familiarization trial* introduced subjects to the apparatus and the apple. The front occluder was flipped down to reveal an apple on the stage in between the relevant boxes (Fig. 1a). The demonstrator looked at the apple and said "now" while the subject's looking was recorded for 10 seconds. In the *second familiarization trial*, the demonstrator reached to one of the three boxes (either the irrelevant box or one of the two relevant boxes based on the condition; Fig. 1b) and said "now" while the subject's looking was recorded for 10 seconds.

All subjects then saw a single *test trial*, which began with a *belief induction event* to establish where the demonstrator believed the apple was. The demonstrator first watched the apple move out of one of the relevant boxes and into the second one. The top occluder was then flipped to block the demonstrator's view of the relevant boxes and the demonstrator stared at the irrelevant box. While the demonstrator's view was blocked, subjects could see the apple move back into the box that it was originally in (Fig. 1c). The demonstrator thus had a false belief about the final location of the apple. The top occluder then flipped back down and the demonstrator reached into the same box that they reached into during the second familiarization trial (Figure 1d). This led to three different test conditions: (1) the demonstrator reached into the irrelevant box (irrelevant reach condition), (2) the demonstrator reached into the box where they last saw the apple (false-belief-consistent reach condition), or (3) the demonstrator reached into the box where the apple was actually located (reality-consistent reach condition). After reaching, the demonstrator said "now" and held his pose for 10 seconds while the subject's gaze was recorded.

As in all subsequent studies, we counterbalanced the box which the apple started in across subjects. Additionally, halfway through data collection, we switched the location of the irrelevant box from the leftmost box to the rightmost.

Video Coding Two coders (blind to condition) measured how long subjects looked during each 10s trial. Inter-rater reliability was high (Pearson's R = .93).

Results and Discussion

We used a linear mixed effects model to predict subjects' log-transformed looking times in a given trial based on the trial type (familiarization 1, familiarization 2, or test), condition (irrelevant reach, false-belief-consistent reach, or reality-consistent reach) and their interaction with random intercepts for each subject. Trial type significantly predicted subjects' looking: Subjects looked significantly less in familiarization 2 (β = -0.25, p = .008) and the test (β = -0.73, p < .001) relative to familiarization 1. Additionally, looking was significantly lower in test (β = -0.48, p < .001) relative to familiarization 2. This pattern of performance suggests that subjects were familiarized as expected before the test events.

Condition and the interaction between condition and trial type were not significant predictors of subjects' looking (see regression model coefficients in Table 1), suggesting that the changes in subjects' looking times were not contingent upon the condition that they saw. Crucially, there were no significant differences in subjects' looking times within the test trial based on which test event they saw.

These results suggest that although subjects across the three conditions were successfully familiarized to the set-up, they did not find any of the reaches in the test trial to be more unexpected than the others. This result replicates a common finding in comparative ToM studies: primates look equally long no matter whether an agent with a false belief behaves in a manner that is consistent with their false belief or consistent with reality. However, this study goes beyond previous work to show that monkeys are also unsurprised when an agent reaches to a completely irrelevant box that never held the object and was physically separated from the other boxes by a barrier. This pattern of results is consistent with the no prediction account of primate false belief performance since subjects' looking responses suggest that no expectation was violated by the reach to the irrelevant box and thus no increase in looking time was observed in the irrelevant reach condition.

One alternative interpretation of these findings, however, is that the subjects *did* form multiple predictions, but those predictions included one in which the demonstrator could have reached into the irrelevant box. In an effort to emphasize that the demonstrator knew that nothing could enter the irrelevant box during the belief induction event, we had the demonstrator stare at the irrelevant box while the apple moved between the two relevant boxes. Unfortunately, this design may have had the opposite effect, causing subjects to predict that the agent might reach into the irrelevant box since he showed interest in that location. Study 2 fixes this issue and makes the "irrelevant box" even more irrelevant to see if monkeys persist in not forming predictions.

Study 2

Methods

Subjects 121 monkeys participated (54 females; mean age = 6.31 ± 4.31). Additional monkeys were excluded: 67 for inattention, 48 for leaving the study area, 8 for displacement by another monkey, 23 for previous participation, 2 for experimental errors, and 6 for approaching the apparatus.

Procedure All subjects saw the same two familiarizations as in Study 1 (Fig. 1a and 1b). However, the belief induction event was changed to heighten the irrelevance of the third box. In this new version, the demonstrator first watched the irrelevant box flip open, revealing that it was empty, and then flip closed again (Fig. 1e). The demonstrator then watched the apple move out of one of the relevant boxes and into the second one. The top occluder then flipped to block the demonstrator's view of the relevant boxes, but this time

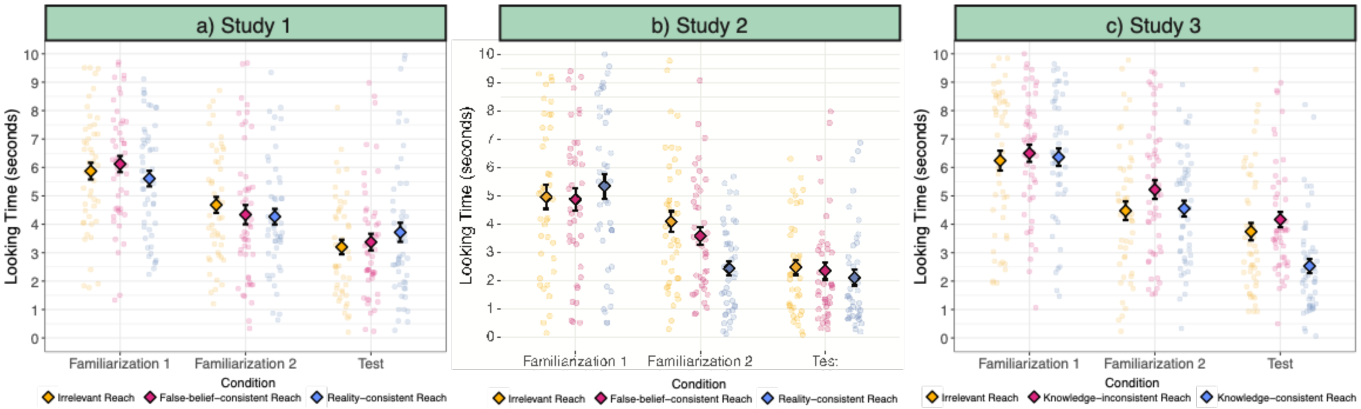

Figure 2: The results of a) Study 1, b) Study 2, and c) Study 3. Figures show subjects' raw looking times (y-axis) across familiarization 1, familiarization 2, and the test trial (x-axis). Dot colors indicate the type of reach seen by subjects in the test trial and darker colored diamonds represent condition averages. While subjects in Studies 1 and 2 did not look significantly longer when a demonstrator with a false belief reached to an irrelevant location, subjects in Study 3 looked significantly longer when a knowledgeable agent reached to the irrelevant location.

stayed upright, parallel to the back of the stage, blocking the demonstrator entirely from the subject's view (Fig. 1e). While the demonstrator's view was blocked, subjects could see as the apple moved back into the box that it was originally in (Fig. 1e). Thus, the demonstrator knew the irrelevant box was empty and had a false belief about the location of the apple.

Once the apple finished moving, the top occluder flipped back down and the demonstrator reached into the same box that they reached into during the second familiarization trial. As in Study 1, this led to three different test conditions: (1) the demonstrator reached into the highly irrelevant box (*irrelevant reach* condition), (2) the demonstrator reached into the box where they last saw the apple (*false-belief-consistent reach* condition), and (3) the demonstrator reached into the box where the apple was actually located (*reality-consistent reach* condition).

Video Coding Coding was the same as in Study 1 with high inter-rater reliability (Pearson's R = .94).

Results and Discussion

As in Study 1, we used a linear mixed effects model to predict log-transformed looking times in a given trial based on the trial type, condition and their interaction with random intercepts for each subject. Trial type significantly predicted subjects' looking time. Specifically, subjects' looking times were significantly lower in the test trial (β = -0.78, p < .001) relative to familiarization 1 and relative to familiarization 2 (β = -0.62, p < .001).

While there was no main effect of condition, there was a significant interaction such that, in the reality-consistent reach condition (see Table 1 for full regression coefficients), looking times decreased significantly more between familiarization 1 and familiarization 2, relative to both the irrelevant reach (β = -0.71, p < .001) and false-belief-consistent reach (β = -0.61, p = .002) conditions. Since subjects across all conditions saw the same familiarizations, this interaction is difficult to interpret, but could indicate that, by chance, monkeys in the reality consistent reach condition habituated to the set-up more quickly. Critically, however, looking times in the irrelevant reach condition did not decrease significantly less (in fact, they decreased significantly more) than the reality-consistent reach (β = -0.43, p = .029) or false-belief-consistent reach condition (β = -0.05, p = .814) between familiarization 2 and the test trial, which is what we would have expected if subjects found the irrelevant reach to be unexpected. Thus, subjects were unsurprised even when the demonstrator reached to a box a truly irrelevant empty box.

Study 2's results are not only consistent with our no prediction account, but also highly inconsistent with a multiple predictions-based account of primates' performance. Given the set-up of the current study, it would be difficult to justify why subjects would have the expectation that the demonstrator would be just as likely to reach to the irrelevant location as the other two locations. Thus, it seems highly unlikely that primates' null performance here can be explained by positing that they maintained three predictions about how the agent would act that each related to one of the boxes.

Study 3

Although Studies 1 and 2 provide evidence in support of the no prediction account of primates' false belief performance, this evidence is in the form of null performance. Unfortunately, it is unclear from the first two studies alone whether primates are demonstrating the posited representational limit of interest or whether they are simply confused by the set-up. Therefore, in Study 3 we test whether primates can successfully make positive predictions when the agent has a true belief in this 3-box set-up.

Methods

Subjects We tested 150 monkeys (72 females; mean age = 5.20 ± 3.28) for Study 3. Additional monkeys were approached but did not complete participation due to inattention (n = 14), leaving the area (n = 33), displacement (n = 12), previous participation (n = 41), approach (n = 2), or due to an experimenter error (n = 2).

Procedure The familiarization trials and condition assignment were identical to Study 1. However, during the test trial, the demonstrator's view was not occluded. Instead, he watched the apple move out of one of the relevant boxes and into the second one and also saw the apple move back into the box that it was originally in (Fig. 1f). Thus, the demonstrator witnessed all of the apple's movements, making him knowledgeable about the apple's final location.

Once the apple finished moving, the demonstrator reached into the same box that he reached into during the second familiarization trial. This led to three different test conditions: (1) the demonstrator reached into the irrelevant box (*irrelevant reach* condition), (2) the demonstrator reached into the box where the apple was not located (*knowledge-inconsistent reach* condition), or (3) the demonstrator reached into the box where the apple was located (*knowledge-consistent reach* condition).

Video Coding Trials were coded as in previous studies with high inter-rater reliability (Pearson's R = .92)

Results and Discussion

Using a linear mixed effects model predicting subjects' log-transformed looking times in a given trial based on the trial type (familiarization 1, familiarization 2, or test), condition (irrelevant reach, knowledge-inconsistent reach, or knowledge-consistent reach) and their interaction with random intercepts for each subject, we found that trial type significantly predicted subjects' looking times: looking times were significantly lower in the familiarization 2 (β = -0.38, p < .001) and the test trial (β = -1.11, p < .001) relative to familiarization 1. Additionally, looking times were significantly lower in the test trial (β = -0.73, p < .001) relative to familiarization 2.

Although condition was not a significant predictor of subjects' log-looking times (see Table 1 for regression model coefficients), we found a significant interaction between condition and trial type such that the decrease in monkeys' looking times between familiarization 2 and the test trial was significantly sharper for monkeys in the knowledge-consistent reach condition relative to the knowledge-inconsistent reach condition (β = -0.52, p < .001) and the irrelevant reach condition (β = -0.57, p < .001). Thus, monkeys spent less time looking when the demonstrator's actions were consistent with his knowledge. However, when the demonstrator reached to the empty or the irrelevant box, monkeys looked longer, suggesting that they found these outcomes more unexpected.

These results suggest that rhesus macaques can form expectations about knowledgeable agents' behavior in our experimental set-up even though they fail to do so when an agent has a false belief. This pattern of performance suggests that primates are not confused by this general experimental set-up and that they instead fail because they are unable to specifically predict the behavior of agents with reality-inconsistent mental states like false beliefs.

Table 1: Regression Coefficients (Studies 1-3)

Predictors	Beta	Standard Error	t-value	p value
a) Study 1				
Trial Type (ref: Familiarization 1)				
Familairization 2	-0.25	0.09	-2.68	0.008
Test	-0.73	0.09	-7.73	< 0.001
Condition (ref: Irrelevant Reach)				
False-belief-consistent Reach	0.04	0.12	0.31	0.758
Reality-consistent Reach	-0.07	0.12	-0.60	0.546
Fam 2: False-belief-consistent Reach	-0.21	0.13	-1.58	0.115
Fam 2: Reality-consistent Reach	-0.08	0.13	-0.62	0.537
Test : False-belief-consistent Reach	-0.01	0.13	-0.10	0.917
Test : Reality-consistent Reach	0.15	0.13	1.16	0.249
b) Study 2				
Trial Type (ref: Familiarization 1)				
Familairization 2	-0.16	0.14	-1.17	0.244
Test	-0.78	0.14	-5.57	< 0.001
Condition (ref: Irrelevant Reach)				
False-belief-consistent Reach	0.01	0.19	0.06	0.951
Reality-consistent Reach	0.09	0.19	0.49	0.625
Fam 2: False-belief-consistent Reach	-0.11	0.20	-0.53	0.594
Fam 2: Reality-consistent Reach	-0.71	0.20	-3.61	< 0.001
Test : False-belief-consistent Reach	-0.06	0.20	-0.30	0.765
Test : Reality-consistent Reach	-0.28	0.20	-1.42	0.158
c) Study 3				
Trial Type (ref: Familiarization 1)				
Familairization 2	-0.38	0.10	-3.80	< 0.001
Test	-1.11	0.10	-11.14	< 0.001
Condition (ref: Knowledge-consistent Reach)				
Knowledge-inconsistent Reach	0.03	0.12	0.24	0.811
Irrelevant Reach	-0.04	0.12	-0.34	0.732
Fam 2: Knowledge-inconsistent Reach	0.12	0.14	0.82	0.413
Fam 2: Irrelevant Reach	-0.05	0.14	-0.32	0.750
Test : Knowledge-inconsistent Reach	0.63	0.14	4.48	< 0.001
Test : Irrelevant Reach	0.53	0.14	3.74	< 0.001

General Discussion

Taken together, this series of studies sheds new light on primates' null performance on false belief tasks. First, we conceptually replicated past studies (Drayton & Santos, 2018; Horschler et al., 2019; Marticorena et al., 2011), finding that monkeys are not surprised when an agent with a false belief behaves in a way that is inconsistent with that belief. However, here we reveal the full extent of this failure by showing that monkeys remain unsurprised even when an agent with a false belief reaches to a completely irrelevant location. We found this pattern not only when the irrelevant location was physically separated from the other possible locations, but also when the agent saw that that location was empty. Finally, we demonstrated that monkeys can make positive predictions about how an agent with a true belief will behave in a closely-matched set-up: when the agent sees the object's final hiding location, monkeys *do* find it unexpected when the agent subsequently reaches to the box that the apple just moved out of and when the agent reaches to an irrelevant box where the apple was never located. Altogether, this pattern of failures reveals a key limit on primates' capacity to predict others' actions: when an agent has a false belief, primates seem to be unable to generate *any predictions* about how they will behave.

Why would primates fail to generate predictions in these cases? One proposed theory (Martin & Santos, 2016; Phillips et al., 2021) is that primates are only capable of representing others' mental states when those mental states are consistent with reality (i.e., representing that the agent knows something that the subject himself also knows). However, once an agent's perspective is out of step with the subject's (e.g., the agent did not see an object switch locations), primates cannot represent that mental state since it has reality-inconsistent content. Our findings not only corroborate this theory, but also provide an additional account of what happens to those representations of others' beliefs once they become inconsistent with reality: namely, those representations are completely dropped and fail to inform any subsequent predictions about the agent's future behavior.

In this way, primates may exhibit a signature cognitive limit when representing the awareness of others. Once primates detect that another agent's perspective is out of step with their own, they may drop all representations of that agent's past knowledge and thus, have no representations on which to base their action predictions, which, in our study, led to equal looking regardless of how the agent acts next. Note that this account nicely explains why subjects in Study 2 fail to predict that the demonstrator should not reach to the irrelevant box even though he saw that it was empty; the subsequent location change (that was unobserved by the demonstrator) may have caused subjects to drop their representation of the demonstrator's knowledge about the contents of the irrelevant box. While additional research is needed to further support this explanation of the behavior observed in our studies, this account is a promising framework for unifying other disparate findings in the study of primate ToM as well.

Crucially, however, the failures that we explore here should not be viewed as an attack on the complexity of nonhuman primate social cognition. A wealth of studies show that primates can make flexible predictions about the behavior of knowledgeable agents across a variety of contexts (Arre et al., 2021; Drayton & Santos, 2017, 2018; Flombaum & Santos, 2005; Hare et al., 2001; Kaminski et al., 2008; Krachun et al., 2009; MacLean & Hare, 2012; Marticorena et al., 2011; Martin & Santos, 2014; Melis et al., 2006; Santos et al., 2006) and our findings do not undermine that remarkable behavioral flexibility. Additionally, primates can use their capacity to represent others' knowledge to perform informative behaviors in order to tell conspecifics about an unobserved threat (Crockford et al., 2012, 2014) and opt to reveal food to a cooperator who can assist them in accessing it (Karg et al., 2015b). Moreover, a perspective-tracking system that does not make any predictions about the behavior of agents with false beliefs may actually be more ecologically useful than one that makes reasonable, but errant predictions (e.g., predictions based on reality).

If primates do not have any expectations about how an agent with a false belief will behave, then how can we rectify this account with some primates' success on anticipatory looking false belief tasks (e.g., Krupenye et al., 2016)? One possibility is that primates' anticipatory looking may fall short of making actual predictions. Under this view, there may be representations capable of directing primates' attention (and thus motivating anticipatory looking), that fail to support the more active predictions needed for success in violation of expectancy experiments. This view would explain why primates sometimes succeed in anticipatory looking false belief studies while showing no prediction in both VOE and other more active competitive false belief tasks (Call & Tomasello, 1999; Drayton & Santos, 2018; Hare et al., 2001; Kaminski et al., 2008; Krachun et al., 2009; Marticorena et al., 2011; Martin & Santos, 2014).

Our task and findings also present new opportunities for exploring the nature of infants' null performance on false belief tasks. Some researchers have argued that infants pass certain types of false belief trials but not others because of contamination from the most recent location of a hidden object (Baillargeon et al., 2018). This proposal is not unlike our own multiple predictions account and could be tested using the experimental set-up we used here. The results of such a study would not only provide important insight into the foundations of ToM in humans, but also into possible differences between infants and primates. On the one hand, it is possible that infants, like primates, also fail to form expectations about the behavior of agents with false beliefs. In this way, human ToM may also begin in a restricted form, representing and predicting others' behavior based only on reality-consistent mental states. This pattern of performance would suggest that the foundations of human ToM are very similar to those of primates. Such findings would also open up new questions regarding the trajectory of humans' ToM development: namely, when do children begin to make predictions about agents with reality inconsistent mental states? In contrast, infants may show a different pattern of performance from primates and instead make multiple predictions about what an agent with a false belief will do next. Under this view, there may be a fundamental difference in infant and primate ToM-capacities – a gulf that may only deepen as human children continue to elaborate upon this more robust social cognitive system.

OSF Repository Available anonymized at https://osf.io/3n8fq/?view_only=a31e511347b543c3b89696796ea335a9

References

Arre, A. M., Stumph, E., & Santos, L. R. (2021). Macaque species with varying social tolerance show no differences in understanding what other agents perceive. *Animal Cognition*.

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited Commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*.

Barone, P., Wenzel, L., Proft, M., & Rakoczy, H. (2022). Do young children track other's beliefs, or merely their perceptual access? An interactive, anticipatory measure of early theory of mind. *Royal Society Open Science*.

Bartsch, K., & Wellman, H. M. (1995). *Children Talk About the Mind*. Oxford University Press.

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbühler, K. (2012). Wild Chimpanzees Inform Ignorant Group Members of Danger. *Current Biology*.

Crockford, C., Wittig, R. M., & Zuberbühler, K. (2017). Vocalizing in chimpanzees is influenced by social-cognitive processes. *Science Advances*.

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*.

Drayton, L. A., & Santos, L. R. (2017). Do rhesus macaques, Macaca mulatta, understand what others know when gaze following? *Animal Behaviour*.

Drayton, L. A., & Santos, L. R. (2018). What do monkeys know about others' knowledge? *Cognition*.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus Monkeys Attribute Perceptions to Others. *Current Biology*.

Garnham, W. A., & Perner, J. (2001). Actions really do speak louder than words—but only implicitly: Young children's understanding of false belief in action. *British Journal of Developmental Psychology*.

Hamlin, K. J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science*.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*.

Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi, K., Kato, S., Hori, Y., Nagai, Y., Iijima, A., Someya, T., & Hasegawa, I. (2020). Macaques Exhibit Implicit Gaze Bias Anticipating Others' False-Belief-Driven Actions via Medial Prefrontal Cortex. *Cell Reports*.

Holland, C., & Phillips, J. (2020). *Proceedings of the Cognitive Science Society*.

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2019). Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. *Cognition*.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*.

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). The goggles experiment: Can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour*.

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*.

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

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*.

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*.

MacLean, E. L., & Hare, B. (2012). Bonobos and chimpanzees infer the target of another's attention. *Animal Behaviour.*

Marticorena, D. C. W., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science.*

Martin, A., & Santos, L. R. (2014). The origins of belief representation: Monkeys fail to automatically represent others' beliefs. *Cognition*.

Martin, A., & Santos, L. R. (2016). What Cognitive Representations Support Primate Theory of Mind? *Trends in Cognitive Sciences*.

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (Pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology*.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*.

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

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*.

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

Santos, L. R., Nissen, A. G., & Ferrugia, J. A. (2006). Rhesus monkeys, Macaca mulatta, know what others can and cannot hear. *Animal Behaviour*.

Scott, R. M., & Baillargeon, R. (2009). Which penguin is This? Attributing False Beliefs About Object Identity at 18 Months. *Child Development*.

Southgate, V., Senju, A., & Csibra, G. (2007). Action Anticipation Through Attribution of False Belief by 2-Year-Olds. *Psychological Science*.

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*.

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development*.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*.

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