Differences Between Human and Non-Human Primate Theory of Mind: Evidence from Computational Modeling

Daniel J. Horschler*[,a], Marlene D. Berke*[,a], Laurie R. Santos[a], Julian Jara-Ettinger[a,b,c]

[a]*Department of Psychology, Yale University*
[b]*Department of Computer Science, Yale University.*
[c]*Wu Tsai Institute, Yale University.*

Abstract

Can non-human primates (NHPs) represent other minds? Answering this question has been historically difficult because primates can fail experimental tasks due to a lack of motivation, or succeed through simpler mechanisms. Here we introduce a computational approach for comparative cognition that enables us to quantitatively test the explanatory power of competing accounts. We formalized a collection of theories of NHP social cognition with varying representational complexity and compared them against data from classical NHP studies, focusing on the ability to determine what others know based on what they see. Our results uncovered that, while the most human-like models of NHP social cognition make perfect qualitative predictions, they predict effect sizes that are too strong to be plausible. Instead, theories of intermediate representational complexity best explained the data. At the same time, we show that it is possible for human-like models to capture non-human primate behavior (NHP), as long as we assume that NHPs rely on these representations only about one third of the time. These results show that, in visual perspective taking tasks, NHPs likely draw upon simpler social representations than humans, either in terms of representational complexity, or in terms of use.

Key words: Comparative Cognition, Computational Modeling, Social Cognition, Theory of Mind

1. Introduction

Like humans, non-human primates have rich social lives: they live in complex social groups, they can act altruistically, they work together to complete everyday tasks, they learn from each other, and they develop lifelong social relations—from dominance hierarchies to close friendships (1–6). At the same

*These authors made an equal contribution and are listed in reverse alphabetical order.

time, their social lives are undeniably simpler relative to that of humans (7, 8): Other primates do not deceive, or persuade each other. They do not invest their time to the transfer knowledge to others. And they do not develop cultures and societies of human-level complexity. What cognitive differences might explain this gap?

Cognitive scientists have long hypothesized that part of the answer lies in our *Theory of Mind*—the ability to represent others' unobservable mental states like their beliefs and desires. Theory of Mind has been implicated in a broad range of human capacities, from language use to moral reasoning (9–11). Its foundational components are at work from early in infancy (12, 13), and many important mental-state inferences are automatic in adults (14, 15). Because of this, questions about the evolutionary origins of Theory of Mind have been studied extensively in non-human primates (see 16–18, for review), but there is surprisingly little consensus about what exactly non-human primates understand about other minds.

Characterizing non-human primate (NHP) Theory of Mind has been historically challenging for many reasons. However, one fundamental challenge has been differentiating between alternative accounts that explain the qualitative data equally well. For instance, consider two competing accounts, one arguing that non-human primates have complex representations of other minds, and a second one suggesting that their behavior can be explained by a simple set of behavioral rules. Suppose further that these two accounts—despite proposing radically different representational contents—converge in predicting that non-human primates should succeed on a given task. If 70% of the tested primates indeed succeed, this is often interpreted as consistent with both accounts (assuming this effect is significantly above chance). Therefore, a task like this one would fail at differentiating between accounts, and this is usually the state of comparative research on Theory of Mind—for every study, there exists a mentalistic and a non-mentalistic account that both qualitatively explain the data (19–22).

This analysis, however, makes two implicit assumptions. The first assumption is that a 70% success rate is equally consistent with the two different accounts. As we show in this paper, this is not always the case: different accounts can predict different effect sizes. This is because, under some representational contents, non-human primates would have a clear representation of others' knowledge and have high confidence about how to react. Under other representational contents, non-human primates might have weak expectations about others' minds, leading to weaker effect sizes in how they react.

The second assumption is that non-human primates were not always *relying* on the posited representations, given the 30% failure rate. Intuitively, however, this failure rate should also inform how much we choose to accept different proposals. For instance, we may be more willing to accept a high failure rate when a theory proposes that solving the task requires complex social inferences, but the same failure rate might raise concerns for a theory that posits a trivial mechanisms that guarantee success.

The challenge of distinguishing between competing accounts on the basis

of qualitative predictions is not unique to comparative cognition. In recent decades, computational cognitive modelling has emerged as a powerful tool to help solve this issue. By formalizing different theories in exact computational terms, it is possible to derive exact effect sizes that can be compared against experimental data. This general approach has proved successful in multiple related fields such as human Theory of Mind (e.g., comparing the explanatory power of mentalistic models versus simple cue-based alternatives; 23), and causal reasoning (evaluating the role of counter-factual reasoning; 24).

Our goal in this paper is to advance a computational methodology for understanding non-human primate social cognition. Because Theory of Mind is a complex cognitive system with many sub-components (including mechanisms for inference, prediction, explanation, and even planning over other minds; Ho et al. 25), here we focused on just one component of ToM that's been well-researched in comparative cognition: the ability to determine what others see and know based on their visual perspective (see 26, for review). The capacity to represent others' perceptual and knowledge states are among the most basic components of ToM, emerging early in human infancy (27), and the question of whether these capacities are shared with our NHP relatives has been the subject of much debate (16, 18, 28–30).

To illustrate this capacity, consider Fig. 1a, which shows a subordinate chimpanzee (the subject; black) and a dominant conspecific (the competitor; gray) on opposite sides of a room with two apples. One apple is at the center of the room and visible to both chimpanzees, while the second apple is behind an opaque barrier and visible only to the subject. If the subject understands that the competitor only knows about the apple in the center of the room, it can use this understanding to strategically decide which apple to go for. Tasks like these were used in now-classical work to show that chimpanzees (*Pan troglodytes*) will preferentially reach for food rewards that are hidden from the conspecific, providing some of the first evidence for visual perspective taking in non-human primates (NHPs), which has since been extended to show more complex forms of perspective taking, including evidence that NHPs know what others can hear (31, 32).

To explore what types of representations underlie this capacity, we developed seven computational models of varying cognitive complexity (Shown in Fig. 1b-h). Each of these models is instantiated as a simple system that is able to complete basic primate experiments implemented in simple two-dimensional grid world. We can therefore use these models to derive exactly how easy or difficult a task should be for non-human primates, according to different models. In addition, this enables us to explicitly manipulate how often non-human primates actually rely on each model's posited representations (hereafter referred to as reliance).

We evaluate our models using three approaches. First we examined each model's qualitative pattern of performance. Each model completed a suite of seminal visual perspective-taking tasks, and we compared each model's qualitative pattern of successes and failures to those documented in comparative studies (Section 3.1). This revealed an ordinal relationship: the more complex

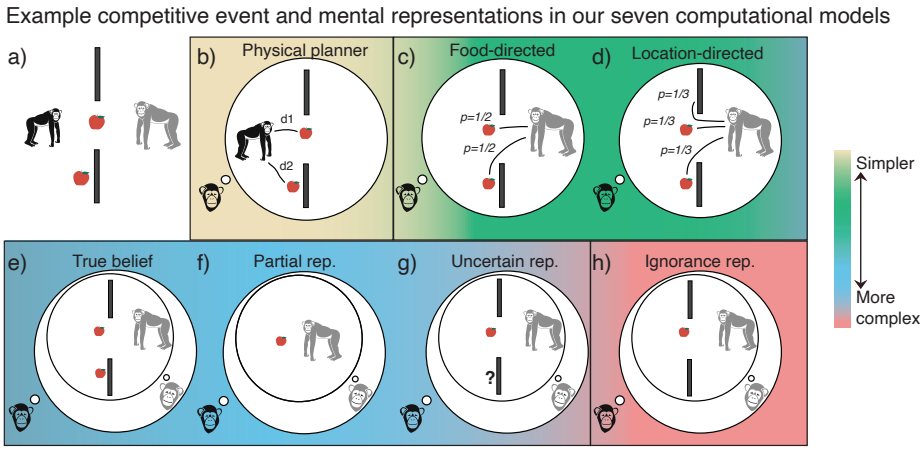

Figure 1: a) Hypothetical competitive event to illustrate the logic of our seven computational models. Here, the subject (small chimpanzee, black) is in front of a dominant conspecific (big chimpanzee, gray) in a room with two apples and two opaque barriers (in black), such that both chimpanzees can see the apple between them, but only the subject sees the apple in front of the barrier. c-h) Posited representations for each our seven computational models, ordered by complexity. Outlined rectangles in black indicate the groupings of models into similar families.

the model, the more it matched the pattern of qualitative results from the empirical studies (i.e., the number of directional success/failure predictors that qualitatively match the empirical studies).

Next, we evaluated each model's capacity to quantitatively replicate the exact effect sizes found in non-human primate (NHP) experiments (Section 3.2). This revealed that comparing theories to black-or-white success/failure behavior omits important information: although the more complex models match the qualitative results, they predict effect sizes that are too strong when compared to empirical data (i.e., they predict that primates should be much better at the task than they actually are). This suggests that models of intermediate complexity better represent non-human primate Theory of Mind.

When a model predicts an effect size that is too strong, this discrepancy can be accommodated by lowering the reliance parameter in our model (i.e., the less that non-human primates rely on the representation, the weaker the effect sizes become). We therefore take this approach to infer how often non-human primates *rely* on the posited representations, for each model to maximize its explanatory power. Through this approach, we can reveal what implicit commitments people must make about reliance when they advance a particular representational theory. In particular, we show that the most human-like models of Theory of Mind can only explain primate behavior by assuming that non-human primates can only access these representations about one third of the time. Therefore, positing that non-human primates (NHPs) have human-like ToM in their representations comes with the implication that the frequency with which NHPs rely on those representations is lower than human's reliance

on mentalistic reprsentations.

2. Experimental paradigms and computational modeling

2.1. Experimental paradigms

Our work focuses on a set of eleven experiments and controls designed to assess non-human primates' (NHPs) understanding of how another agent's visual perspective affects their knowledge (Table 1). These eleven experiments fall into five general paradigms (shown in Fig. 2) which follow a similar structure: A subject and a dominant competitor face one another on opposite sides of an enclosure with two food rewards that they ultimately compete for. Because dominant individuals in NHP species tend to monopolize food and respond antagonistically to challenges from lower-ranking individuals (33), these paradigms create a pressure for the subordinate subject to exploit any privileged knowledge of food locations to ensure they obtain a reward.

In the Center-Wall paradigm, the subject and the dominant conspecific face each other with a visible food reward directly between them, but the subject can see a second food reward that is hidden from the competitor's view by a barrier. The Open-Hidden paradigm is similar, with the difference that the two rewards are equidistant from the subject. In both of these paradigms, NHPs are significantly more likely to take the hidden reward. In the Transparent-Hidden Routes paradigm, the subject can reach for a food reward through a left or a right path, but one of the paths is hidden from the competitor's sight for longer than the other. Here, NHPs are more likely to reach through the hidden route than through the transparent one.

The final two paradigms are control conditions. In the Hidden-Hidden paradigm both food rewards are visible to the subject but hidden from the competitor's perspective. In the Open-Transparent paradigm, the situation is identical to the Open-Hidden paradigm with the difference being that the barrier is transparent (and both the subject and competitor therefore see both food rewards). In both of these paradigms, NHPs show no systematic preference for either food reward.

2.2. Computational models

To understand the relationship between a theory's complexity and its explanatory power, we implemented seven computational models that vary in representational richness. These models are not meant to be accurate and faithful representations of existing theories. Only the proposers of these theories can specify what exact computational implementation would capture their nuanced positions. Instead, the range of theories that we present are a representative sample from the space of possible theories, varying across levels of complexity. These theories fall broadly in four families: *Egocentric*, *Behavioral*, *Mentalistic*, and *Full Theory of Mind*. For clarity, we explain each model in the context of the hypothetical event shown in Fig. 1a (a more technical description of the models is available in Methods and Materials). We begin by explaining each

Paradigm	**Paper**	**Exp**
Center-Wall	Hare et al. (34)	E1
Open-Hidden	Bräuer et al. (35)	E2
	Canteloup et al. (36)	E1
	Hare et al. (34)	E2-E4
Transparent-Hidden Routes	Hare et al. (37)	E2
	Melis et al. (32)	E1
Hidden-Hidden	Hare et al. (34)	E3-E4
Open-Transparent	Hare et al. (34)	E5

Table 1: Papers and experiments used to evaluate our computational models. Note that Hare et al. (34) Exp 3 and 4 had multiple conditions that use more than one paradigm and are therefore repeated. All experiments tested chimpanzees (*Pan troglodytes*) except for Canteloup et al. (36), which tested Tonkean macaques (*Macaca tonkeana*).

model's posited mental representation, and we then explain how we integrate the notion of reliance into the models (i.e., the ability to modulate how often non-human primates actually use the posited representations).

The *Egocentric* family is the simplest one, consisting of theories where NHPs entirely ignore the competitor. We implemented only one model in this family: The *Physical planner* model, where NHPs simply attempt to go towards the food reward reachable by the shortest path (Fig. 1b). Next, the *Behavioral* family consists of proposals where NHPs form non-mentalistic expectations about competitor behavior, and react accordingly. In the *Food-directed behavior* model, the NHP expects the competitor to pursue one of the food rewards (possibly learned from experience seeing conspecifics regularly go towards food), but lacks a mechanism for predicting which one, therefore placing an equal probability over each of the food rewards (Fig. 1c). The *Location-directed behavior* model extends the behavioral expectations to an expectation that competitors might also check locations they cannot see (which could also be learned from experience observing conspecifics without necessarily representing epistemic states). This model therefore places an equal probability distribution over all the food rewards and places where a food reward could be (even if none is there; Fig. 1d).

The *Mentalistic* family consists of models that represent the competitor's mind, but lack full human-like mental state representations. These models are loosely inspired by proposals that NHPs do represent others' mental states, but in a markedly limited way (e.g., Martin and Santos 38). The *True belief* model captures the idea that NHPs attribute their own knowledge to others and expect competitors to act based on this knowledge (Fig. 1e). Although the subject and competitor share the same representation of the world, they may still pursue different food rewards (e.g., each individual might go for its closest food reward, which could be different for the subject and the competitor). In the *Partial representation* model, the NHP can compute what parts of the environment

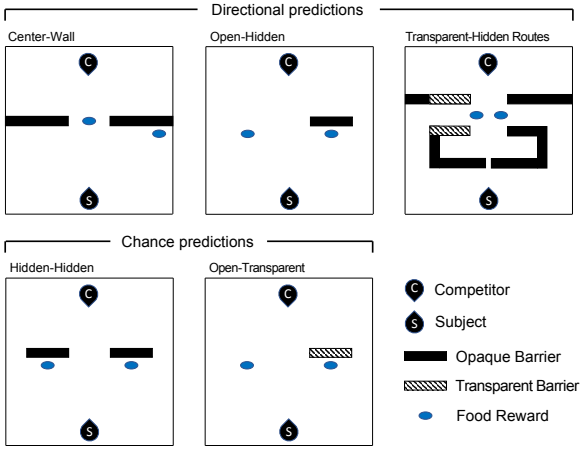

Figure 2: Depictions of the experimental set-up for each of the five paradigms. The top row shows paradigms with directional predictions: if non-human primates represent others' knowledge, they should preferentially go towards the reward (or take the route) that is hidden from the competitor. The bottom row shows paradigms with chance predictions: when the two rewards are simultaneously known or hidden from the competitor, the subject should show no reliable preference for which one to pursue.

are in the shared visual field, and only represents those in a conspecific's mind. This model therefore has no representation of the competitor's awareness (or lack thereof) about objects or locations hidden from the competitor's view (Fig. 1f). That is, any objects outside of the conspecific's visual field are simply not represented.

Finally, in the *Uncertain representations* model, the NHP assumes shared knowledge for objects that are in visual common ground, but is uncertain whether the competitor is or is not aware of objects outside the competitor's field of view. This model differs from the *partial representation* model in that the subject is thinking about the possibility that the conspecific might know about the hidden reward (but is at the same time, unsure about whether they know about it or not). This model therefore builds predictions by integrating the two epistemic hypotheses about the competitor (knows or does not know about the hidden object, using a uniform prior; Fig. 1g).

Finally, the *Full ToM* family consists of theories of human-like Theory of Mind. Although different researchers have proposed different theories of what full human-like ToM consists of (e.g., Hutto 39, Gordon 40, Jara-Ettinger 41), these theories only make different predictions in complex cases that go beyond those captured in the paradigms considered here. For this family we therefore include only an *Ignorance representation* model, where NHPs have an accurate representation of their competitor's visual perspective which both represents their ignorance of hidden object and their knowledge of visible objects (Fig. 1h).

2.2.1. Modeling reliance

Each computational model can generate expectations about how non-human primates should behave, if they were always relying on the posited representations (note that this doesn't imply full success; if the model is uncertain about what the conspecific knows, their behavior will reflect this). We then extended each model to include a *reliance* parameter, which represents how often an NHP uses the model's representations to determine how to act. This reliance parameter therefore captures the possibility that NHPs might fail to use their social representations for a variety of reasons such as a lack of motivation, the cognitive cost of using their ToM, a lack of trust in their predictions, distraction, not caring about the competitor, a failure to inhibit default behavior, or general noise in their choice behavior. Formally, if a model indicates that the subject should pursue reward A with probability $p_{\text{model}}(A)$, then the subject's final probability of pursuing food A ($p_{\text{subject}}(A)$) is given by

$$p_{\text{subject}}(A) = rp_{\text{model}}(A) + (1 - r)p_{\text{egocentric}}(A) \tag{1}$$

where $p_{\text{egocentric}}(A)$ is the probability that the NHP would pursue reward A when they fail to to consider the competitor (i.e., fail to use the model's representations). This term is determined by what the NHP would do if the competitor wasn't there (therefore fully ignoring them). When reliance $r = 1$, NHPs always use the posited representations. Conversely, when $r = 0$, NHPs are entirely unable or unwilling to use the social representation. Intermediate reliance values indicate that NHPs inconsistently rely on the model when deciding what to do.

3. Results

3.1. Qualitative model performance

We first evaluated each model's capacity to replicate the qualitative pattern of successes and failures documented in the NHP literature. For this test we estimated the reliance directly from the NHP experimental data, obtaining a reliance of $r = 0.6$ that we fixed for this analysis (i.e., assuming that these tasks elicit their social representations 60% of the time; see Sec. 5.4.1). Because each computational model predicts a continuous effect size (rather than an absolute success or failure), we coded model behavior as predicting chance performance when its preference for both food rewards was below $t = 0.6$ (see Section 5.3.1 for details), and as making a directional prediction otherwise. Altogether, this first qualitative analysis used reliance $r = 0.6$ and threshold $t = 0.6$.

Figure 3a shows the percentage of paradigms that each model replicates, revealing an ordinal relationship where the models with the most complex social representations best replicated the qualitative experimental record. Figure 3b shows each model's exact performance on each paradigm (which can also be interpreted as the model's predicted success rate). A detailed walk-through of each model's qualitative behavior is available in Supplemental Materials Section 1.1. A robustness analysis revealed that these results are representative of

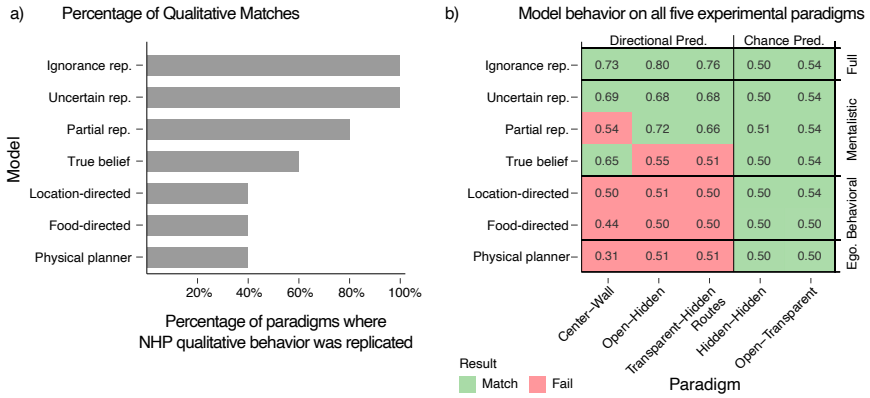

Figure 3: Results from the qualitative analysis. a) Percentage of paradigms where each model matches the qualitative pattern found in NHP literature (with threshold $t = 0.6$). Models (y-axis) are ordered by complexity with most complex representations at the top. This figure reveals an ordinal relation where the most complex models better replicate the full qualitative pattern of NHP behavior. b) Detailed results showing each model's predicted effect sizes (i.e., probability of success, from 0 to 1) for each paradigm. Each row represents a computational model and each column one of the five paradigms. The first three paradigms have shown preferential choices in NHPs while the last two paradigms have documented chance performance. Background color indicates whether the model prediction is consistent with NHP behavior or not, at the $t = 0.6$ threshold (such that panel a is showing the percentage of green cells in this figure).

general model performance under different threshold values t (see Figs. S1-S2 in Supplemental Materials).

3.2. Quantitative model performance

Our qualitative analyses suggest an ordinal relationship between model complexity and consistency with NHP behavior. However, each model makes exact predictions about expected effect sizes. As a consequence, even if a model makes the correct directional prediction, it may lack explanatory power if it either over- or under-predicts an effect size relative to observed NHP behavior. To test this, we calculated the exact probability of each model generating the empirical pattern of NHP choices ($P(Data|Model)$), using the seven experiments that reported enough quantitative information (Exact values available in Fig. S3 in Supplemental; see Section 5.4 for details).

To interpret these results, we performed pair-wise model comparisons, testing each model's relative ability to explain the data using Bayes factors with a uniform prior (i.e., assuming all models were all a prior equally likely; see Section 5.2.2). Fig. 4 shows the results from this analysis. In contrast to the qualitative analyses, the *Uncertain representation* model now outperformed all other models: the data from every single paradigm supported this model over 100 times more than any other model (i.e., $BF > 100$ for all cases). Interestingly, the other two *Mentalistic* models (*Partial representation* and *True belief*) also outperformed the most complex model (*Ignorance representation*). This

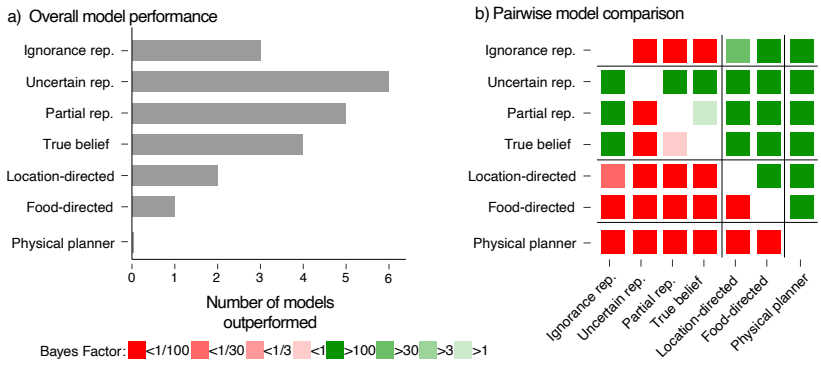

Figure 4: a) Number of competing models that each model outperforms (as determined by Bayes Factors). The y-axis shows our seven computational models ordered by cognitive complexity (most complex at the top), and the x-axis shows number of competing models that it outperforms. The three *Mentalistic* models were the highest ranked, with the *Uncertain representation* model beating all other models. The most complex model —*Ignorance representation* — performed worse than the *Mentalistic* models, and the *Behavioral* and *Egocentric* models performed the worst. b) Model comparison matrix with Bayes Factors. Each cell represents how many times more probable the model on the y axis is than the model on the x axis (with y axis using the same order as the first panel). Green indicates Bayes Factors higher than 1 (i.e., row model outperforms column model; panel a therefore shows the sum of green cells in the row), and red indicates Bayes Factors below 1 (i.e., column model outperforms row model).

is because *Ignorance representation* predicted effects that were much stronger than the observed ones. Finally, models in the *Egocentric* and *Behavioral* families the poorest, because they predictions were outright incorrect (as already revealed in the qualitative analysis).

These results suggest that, under a quantitative analysis, the curvature between model complexity and explanatory power has an inverted U-shape: simple models predict effects that are too weak, while the most complex model predicts effects that are too strong. Models of intermediate complexity best explained the effect sizes observed in the NHP experiments.

3.3. Estimating model-posited reliance

So far, our model evaluation assumed that NHPs rely on their social representations 60% of the time—an estimate derived directly from the empirical data (Section 5.4.1). However, our computational models allow us to take a different approach: searching for what reliance parameter maximizes each model's potential to explain NHP data.

Figure 5 shows each model's explanatory power as a function of reliance. The *Physical planner* model is insensitive to the reliance parameter because, by design, it does not have any representation of the competitor's behavior. The *Behavioral* models (*Food-directed behavior* and *Location-directed behavior*) and the *True belief* model showed a similar structure: the stronger the assumption

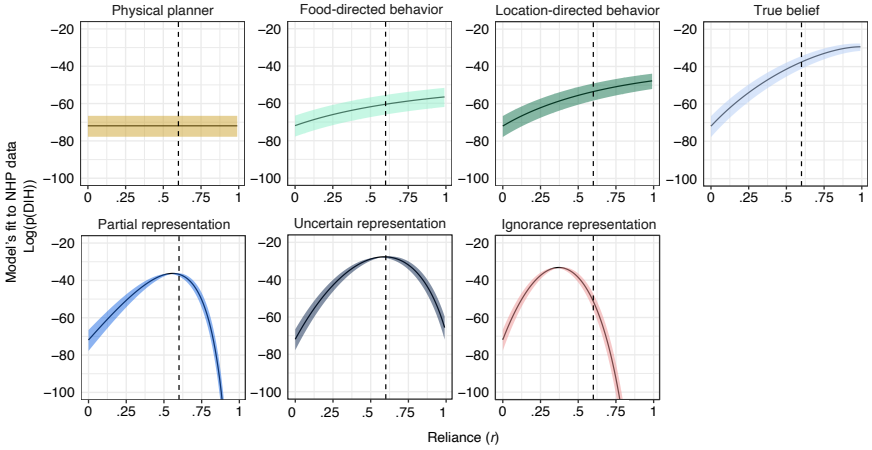

Figure 5: Model explanatory power as a function of the hypothesized reliance r. The x-axis shows reliance. When $r = 0$, the subjects never rely on the posited social representations and when $r = 1$, the subjects always rely on the model's posited social representations. The y-axis shows the probability (in log-space) of the model replicating the experimental data under that reliance value. The dashed vertical line represents the reliance of $r = 0.6$ estimated from data. All models produce the same likelihood (LL=−77.7) at $r = 0$ because at this value they all simply express an egocentric planner.

that primates always use these representations (high r), the better these models can explain the empirical data. The other two models in the *Mentalistic* family (*Uncertain representation* and *Partial representation*) were best able to explain the empirical data under the assumption that subjects rely on their social representations more than half of the time, but critically, not all of the time. Finally, the *Ignorance representation* model's ability to explain the empirical data rapidly decreased when it was posited that subjects always relied on their underlying social representations, because the model expected larger effect sizes in the empirical data.

This analysis enabled us to calculate what reliance parameter maximizes each model's explanatory power (i.e., the r value at which each curve in Fig. 5 peaks). The results are shown in Fig. 6a (with Fig. 6b showing the posterior probability over reliance, as estimated using the NHP data, and the vertical line representing the expected reliance $r = 0.5$; see Sec. 5.4.1). The *Egocentric* and *Behavioral* models (in yellow and green) continued to show the worst performance. By contrast, the *Mentalistic* models (*Partial representation*, *True belief*, and *Uncertain representation*) and the *Full* model (*Ignorace representation*, all achieved high explanatory power (high value on the y-axis), but by positing different degrees of reliance. This reveals that, in principle, models in any of these two families can explain NHP behavior in these perspective-taking tasks, but each model carries a different commitment to how often NHPs rely on the posited representation. If NHPs have human-like *Ignorance representations*, then they must be relying on them very infrequently; only 37% of time (i.e.,

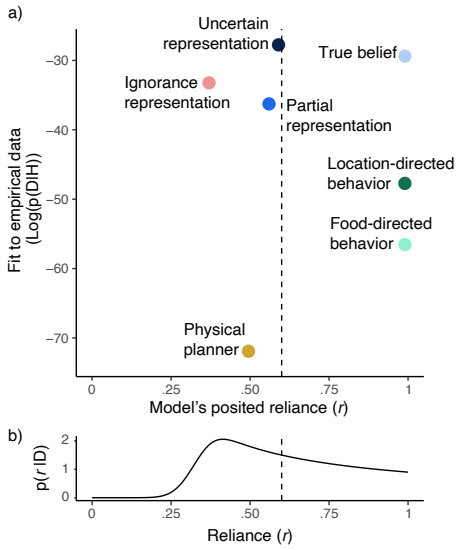

Figure 6: a) Degree of reliance (r) (x-axis) that maximizes each model's explanatory power (y-axis). b) Posterior distribution over NHP reliance r inferred from experimental data. Vertical dashed line shows expected reliance $r = 0.6$ estimated from the empirical data.

$r = 0.37$). This is intuitively consistent with the fact that, if an adult human were placed in this experimental context, one might expect their performance to be much stronger than those of chimpanzees and macaques.

On the other end, the *True belief* model best explains the data under the assumption that NHPs always rely on their ToM in these tasks, with $r = 1.0$. That is, theories that posit that NHPs simply attribute their own knowledge to others best explain NHP data under the assumption that this attribution happens all the time in these tasks, such that NHP behavior is never influenced by task demands, attentional lapses, motivational changes, or other extraneous noise (an intuitively unlikely proposition).

Finally, *Partial representation* and *Uncertain representation* best explain NHP behavior under moderate ToM reliance, with $r = 0.56$ and $r = 0.60$, respectively. Interestingly, the *Uncertain representation* model reached the highest explanatory power from all models and, surprisingly, did so with a reliance parameter $r = 0.6$ that matched our estimate from the empirical data.

4. Discussion

Here, we sought to evaluate non-human primate (NHP) Theory of Mind (ToM) through computational modeling, using their performance on classic visual perspective-taking tasks as a case study. Formalizing theories of NHP social cognition as computational models revealed three broad conclusions. First, when evaluating the *qualitative* pattern of successes, we found an ordinal re-

lationship where increasing a model's representational complexity resulted in more NHP-like patterns of behavior. This is broadly consistent with standard interpretations of this work, suggesting that NHPs have a human-like capacity for visual perspective taking. However, each computational model predicted different effect sizes. Our second analyses therefore revealed that, when evaluating the *quantitative* model fit to NHP behavior, *Mentalistic* models of intermediate complexity best explained the data. This is because, although the most human-like model predicts all of the qualitative effects documented across the NHP data we considered, it also predicts effect sizes that are too strong compared to the experimental record.

Our approach further allowed us to estimate how often NHPs *rely* on the posited representations, according to different models—a commonly unacknowledged component of theories. This revealed two types of proposals with comparable explanatory power. The first proposal is that NHPs have human-like visual-perspective taking capacities, but use them less than 40% of the time in these paradigms. The second proposal is that NHPs rely on simpler cognitive representations and use them around 60% of the time.

This tradeoff between model complexity and reliance is of particular relevance to the question of whether NHPs have a human-like ToM. Human ToM is characterized not only by its cognitive richness, but also by how often we use it in a wide range of contexts, and sometimes even automatically (14). For instance, people will sometimes automatically track each other's knowledge from visual perspective (42) and make automatic common ground inferences (15). Our work therefore suggest that, even if NHPs have a ToM of similar complexity to humans (in the context of visual perspective-taking), then its usage is surprisingly limited compared to how readily human adults use their ToM in visual perspective-taking tasks. This is particularly interesting given that the experimental paradigms considered in this paper were designed with ecological validity in mind, using competitive situations with known conspecifics. In short, if NHPs have human-like social representations, then they are not applying them with human-like frequency.

Methodologically, our work also provides a general framework for computational comparative cognition. Our work focused on seven computational models that vary in degree of cognitive complexity, but they were not intended to cover the full space of conceptual theories that people have. Other researchers and theorists can use our framework to test the performance of their own theories against the models that we present here. Thus, our work contributes a set of benchmarks for theory performance and a framework for computationally implementing and testing the ability of verbal theories to explain empirical data.

One limitation of our work is that our model's *reliance* construct is composed at least two underlying factors. The first factor is the cognitive effort associated with using the representations. That is, reliance can be low because NHPs find using their social representations to be effortful, and this cognitive cost might be exacerbated in increasingly complex paradigms. The second factor is NHPs' motivation to rely on social representation. In other words, reliance also captures how much an NHP might care about participating in the task.

While our approach was able to extract how much reliance different theories implicitly posit, it leaves open the question of what parts of NHP reliance are due to cognitive demands, motivation, or other factors.

While our work focused specifically on visual perspective-taking, ToM has many other components which have been studied in non-human primates, including gaze-following (43–47), awareness representations based on an agent's past perceptions (31, 48–51), and belief representations (38, 52–56). Our work offers a general methodological approach that shows how experimental paradigms can be standardized and placed in a common representational framework that allows for the design and evaluation of computational models. Doing so, however, will require extending this approach to capture paradigm-specific reliance (i.e., some experimental paradigms might increase NHP's reliance on their available social representations; e.g., (53)), and possibly even model switching (i.e., NHPs might use different representations in different tasks). Even within visual perspective-taking tasks, some variations in the paradigms appear to reduce visual perspective-taking success (57). These failures have been interpreted as revealing that NHP social cognition is most visible in more ecologically-relevant competitive contexts (48), and where food reward are far enough apart that they enhance attention to physical costs. We hope that our framework will allow for the effects of these paradigm-specific differences on NHP behavior to be further tested computationally.

5. Methods

5.1. Computational framework

5.1.1. General computational representation

Our general framework builds on Markov Decision Processes (MDPs) (58). MDPs are a general framework for modeling how agents take sequences of actions in an environment to obtain rewards. While MDPs are often used to model first-person behavior, work in human Theory of Mind has shown that they can also be used to model expectations about how other agents will act to obtain rewards, therefore serving as a framework for Theory of Mind (23, 59, 60). Critically, MDPs model behavior through an assumption that agents move efficiently towards their goals—an assumption that both human and non-human primates share (61, 62). However, because classical MDPs always generate only optimal plans, here we instead use probabilistic MDPs. Probabilistic MDPs create graded expectations over behavior, allowing for the possibility that agents will make errors in decision making and planning, particularly when the value of different action plans is similar. An extended presentation of probabilistic MDPs can be found in (59).

Computational models were implemented using the Bishop software package (https://github.com/julianje/Bishop). Each paradigm was modeled as a gridworld with barriers, food rewards, and an agent. The agent could move in the four cardinal directions as well as diagonal directions: up, down, left, right, up-left, up-right, down-left, or down-right. Each movement incurred a small

cost of 0.25 and each food reward had a value of 50. The small cost induces an expectation for efficient action (making shorter action paths less costly), and the reward was set to be large enough such that it would always be higher than the costs (i.e., to avoid situations where the food was not rewarding enough to justify the cost of getting it; as this has never been observed in the experimental paradigms we considered). Note that exact reward values have important consequences in environments where different objects have different rewards, but this was not the case in our paradigms, where all food rewards are always identical.

As the utility of two competing food rewards becomes more similar, it should be harder for an organism to reliably identify which ones is better (e.g., it is easy to tell that it's better to get a food reward that is one step away from an identical food reward that is fifty steps away; but it is harder to do this if the difference were, say, thirty-three vs thirty-four steps in different directions). To account for this, we used the standard approach of softmaxing the utility functions (23, 59, 63, 64), using a temperature parameter of $\tau_{\text{choice}} = 1$. During action planning (i.e., how agents move in space, rather than how they make choices), we also introduced a small probability of errors by softmaxing action plans with temperature $\tau_{\text{action}} = 0.01$. Note that, while the exact success probabilities that each model produces are affected by softmax, the model rankings in our analyses do not change because the same parameter was applied to all models. The details on how each paradigm was implemented are available in Supplemental Section 2.

5.1.2. Model implementation

All models were implemented through Markov Decision processes (MDPs). All MDP environment representations are available in our OSF repository. In most paradigms, we use MDPs to model how the subject represents the competitor's behavior, with two exceptions. The first exception is the physical planner, where the MDP represents the subject's own planning module (as they are acting egocentrically).

The second exception is the Transparent-Hidden routes paradigm. In this paradigm, the subject must consider which of two routes is most likely to be detected by the competitor. Therefore, the MDP represents how the competitor will reason about the subject's movements (rather than representing how the competitor moves towards food rewards). To achieve this, each model attributes a mental representation to the competitor, and then uses this attributed representation to track how the competitor will reason about the subject's movements. For each action plan (i.e., route the subject could take), the model then integrated a probability that the competitor would detect the subject, and the probability of obtaining a reward was set as the probability of not being detected. For most accounts, we used 0.9 as the probability of detection so as to account for the possibility that the competitor might not be able to react on time, or notice the movement.[1]

[1] While different detection probabilities would change model predictions, this would only

Egocentric models

Physical planner. In the physical planner model, each paradigm map was initialized as a single MDP with the agent in the subject's starting position. All food rewards and barriers were present on the map. This approach therefore modeled how the subject would act if it did not represent the competitor at all.

Behavioral models

Food-directed behavior. In the food-directed behavior model, each paradigm map was initialized as two MDPs, each with the agent in the competitor's starting position. Only one of the food rewards was present in each MDP, but all barriers were present in both MDPs. Together, these MDPs therefore generated the expected behavior of the competitor if it were going to each food reward, and predictions were subsequently averaged across MDPs (with equal weighting). This approach modeled an expectation that the competitor would choose one of the food rewards, but lacked a mechanism for predicting which option the competitor may favor, therefore placing an equal probability on each location or route. Because this model specifies a behavioral expectation that competitors pursue food rewards (but no mechanism for reasoning about how they detect others), the probability of detection in the Transparent-Hidden paradigm was set to 0.

Location-directed behavior. In the location-directed behavior model, each paradigm map was initialized in a similar way as the *Food-directed behavior* model, but with the addition of MDPs including a hypothetical food reward in each location unobservable to the competitor (e.g., behind the wall that never actually hid a food reward from the competitor in the Center-Wall paradigm). This approach therefore modeled an expectation that the competitor may search by considering all possible combinations of the locations where the competitor might see or look for food (including areas behind barriers where no food is currently present). Because this model expresses the idea of learned behavioral patterns, in the Transparent-Hidden Routes we apply this principle to detections, such that the subject has general expectations that their behavior is sometimes detected and sometimes is not. Therefore, this expectation applied to the detection probabilities assigned to each route rather than to the food reward locations (i.e., considering all possible combinations of 0.9 and 0 detection probabilities for opaque and transparent routes).

Mentalistic.

True belief. In the true belief model, each paradigm map was initialized as a single MDP with the agent in the competitor's starting position. All food rewards and barriers were present on the map. This approach therefore modeled a true belief default, where the subject attributes their own knowledge to competitors, and uses this representation to predict their behavior (note that

affect one of the five paradigms, and only some of the models for that paradigm. As Figure S3 in supplemental materials shows, the two models that best quantitatively capture behavior in the Transparent-Hidden paradigm (*True belief* and *Physical planner*) do not depend on this parameter. So our overall results (favoring *Uncertainty representation*) do not hinge upon this parameter setting.

this representation does not imply that both agents will pursue the same reward; e.g., each agent might go for the food reward closest to them). In the Transparent-Hidden routes paradigm, detection probabilities for both routes were set to 1, in accordance with the idea that the subject always represented that the competitor shared the subject's knowledge (including their location).

Partial representation. In the partial representation model, each paradigm map was initialized as a single MDP with the agent in the competitor's starting position. However, this MDP excluded areas from the paradigm map which were occluded from the competitor's view as well as the barriers and food rewards on the parts of the paradigm map occluded from the competitor's view (essentially cutting out pieces of the map; map files available in OSF repository). This approach therefore modeled the possibility that the subject was entirely unable to represent the competitor's awareness about objects or the content of locations hidden from their view using a limited representation that only considered regions that were visible to both agents. In the Transparent-Hidden routes paradigm, the transparent route had a detection probability of 0.9, since the subject expects the competitor to share information in common ground. The hidden route was not represented in the map and therefore had no associated detection probability.

Uncertain representation. In the uncertain representation model, each paradigm map was initialized as MDP(s) with the agent in the competitor's starting position. In each MDP, all food rewards that were visible to both the competitor and subject were included on the map, as were all barriers. In paradigms that included food rewards visible to the subject but not the competitor, we included one MDP with these food rewards present on the map and another MDP with them absent from the map, and predictions were subsequently averaged across MDPs. In the Transparent-Hidden Routes paradigm, this approach applied to the detection probabilities assigned to each route, such that the detection probability applied to hidden routes (0.45) was half of that applied to transparent routes (0.9). This approach therefore modeled the possibility that the subject assumed common knowledge for objects and routes that were in plain sight for the subject and competitor, but had uncertainty about whether the competitor was aware of any hidden object or route. This model therefore built predictions by integrating the two epistemic hypotheses about the competitor (knows or does not know about the hidden object or route, using a uniform prior).

Full.

Ignorance representation. In the ignorance representation model, each paradigm map was initialized as a single MDP with the agent in the competitor's starting position. All barriers were present on the map, but only food rewards visible to the competitor were present. In the Transparent-Hidden Routes paradigm, the detection probability applied to hidden routes was 0, while the detection probability applied to transparent routes was 0.9. This approach therefore modeled human-like ToM in theses tasks, in which the subject has a complete representation of the competitor's knowledge and ignorance based on the competitor's visual perspective, and uses it to predict their actions accordingly.

5.1.3. Integrating social predictions into decision making

So far, the models in the *Behavioral*, *Mentalistic*, and *Full ToM* families describe different theories of how NHPs might expect their competitor to behave—expressed as a probability distribution over action plans—but they do not yet describe the subject's own behavior. To transform expectations about the competitor's behavior into first-person decisions about how to act, we formalized the probability of the subject choosing food reward A as

$$p_{\text{model}}(A) = 1 - p_C(A) \tag{2}$$

where $p_C(A)$ is the probability that the competitor will also pursue food reward A. This estimate, $p_{\text{model}}(A)$ was then combined with reliance as described in the main text.

5.2. Model evaluation

5.2.1. Model predictions

To generate model predictions, we first estimated the choice probabilities for each MDP's probabilities (i.e., the probability of choosing each of the two food rewards). We achieved this via Monte Carlo sampling, using 5000 samples per MDP, such that an MDP's probability of choosing a food item equals the proportion of times that this reward was chosen in the simulations. Final model predictions were then obtained by combining the probabilities of different MDPs as specified by the model description (e.g., averaging between the MDPs in cases where the subject holds multiple hypotheses).

5.2.2. Quantitative model comparison

Our quantitative analysis directly compared model performance through Bayes factors. For each pair of models M_1 and M_2 we calculated:

$$\text{BF} = \frac{p(D|M_1)p(M_1)/p(D)}{p(D|M_2)p(M_2)/p(D)} \tag{3}$$

We assumed a uniform prior distribution over models, such that the Bayes Factor becomes the likelihood ratio.

5.3. Parameter settings

5.3.1. Significance threshold in qualitative analysis

Our first analysis required setting a threshhold value above which to consider an effect directional, and below which to consider NHPs to show no preference. In the relevant literature, effects are considered to be directional when they significantly differ from chance (0.5). But how much deviation from 0.5 is significant depends on the sample size.

In the experiments considered, there was enough power to detect an effect of around 0.6. Therefore, we set the threshold value $t = 0.6$. For a sense of typical effect sizes found in the behavioral experiments under consideration: NHPs chose correctly ˜80% of the time in experiments with some of the very

strongest effects (e.g., Hare et al. 34, E1), but more typically oscillated around 60-70% correct (e.g., Hare et al. 34, E3; Bräuer et al. 35, E2) or even slightly below (e.g., Hare et al. 34, E4; Melis et al. 32, E1).

5.4. Behavioral data

To quantitatively evaluate our models, we compared model predictions to effect sizes calculated from the data reported by the experiments in Table 1. For (35), (37), and the Hidden-Hidden conditions of (34) E3 and E4, data on the proportions of choices were incomplete or unavailable, and thus we were unable to include these experiments in our quantitative model evaluations. For most other studies, the proportions of choices were clearly reported. The experiments for which we could most easily recover the choice data from the reported proportions and number of trials were (34) Experiments 1 (45 choices of the hidden food out of 54 trials), E2 (20 out of 27), E3 Open-Hidden condition (52 out of 83), E4 Open-Hidden condition (62 out of 108), and (36) Experiment 1, Condition 2 (95 out of 125). Choice data for (34) E5 and (32) E1 were also recoverable, but required making a few reasonable assumptions. In (34) E5, we assumed a total of 11 subjects with 12 trials per subject based on indirect details in the paper, enabling the recovery of choice data (73 out of 132). For (32) E1, by assuming that trials in which a subject did not make a choice were excluded and by reading mean choices from Figure 2, we recovered the choice data (72 out of 126). When two versions of the same experiment were available that differed only in the timing of the competitor's release (Canteloup et al. 36, E1; Hare et al. 34, E5), we used the version with a longer delay, which better controls for the possibility that the subject's choice could be influenced by first observing the competitor's direction of movement. A summary of this data is shown in Table S1.

5.4.1. Estimating reliance from data

Our second analysis required estimating the reliance parameter r. However, this value is not directly observable, since the observed NHP choices reflect a combination of signal from the primate's unknown ToM ($p_{model}(A)$) and egocentric behavior when subjects do not choose according to their ToM. Conveniently, egocentric behavior predicts chance performance ($p_{egocentric}(A) = 0.5$) in most paradigms, meaning that the reliance parameter r can be thought off as a weight that trades off signal (behavior reflecting ToM) and uniform noise (since egocentric behavior predicts chance performance). In these cases, predictions about the NHP's choice (Eq. 1) become:

$$p_{\text{subject}}(A) = r p_{\text{model}}(A) + (1 - r)(0.5) \tag{4}$$

To obtain a model-agnostic estimate of r (reliance), we performed joint inference over $p_{\text{model}}(A)$ (rather than deriving it through our models) and r, conditioned on $p_{\text{subject}}(A)$ (i.e., the subject's probability of choosing reward A).

The more observed choice data available, the more accurately the latents $p_{\text{model}}(A)$ and r could be inferred. We assumed that every experiment from the

same paradigm should have the same underlying $p_{\text{subject}}(A)$ value. We therefore used the Open-Hidden paradigm, which had the most data available. Thus, the data used to estimate this parameter were from (34) Experiments 2 (20 choices of the hidden food out of 27 trials), E3 (52 out of 83), E4 (62 out of 108), and (36) Experiment 1, Condition 2 (95 out of 125).

The most direct way to perform this inference would be to use a point estimate for $p_{\text{subject}}(A)$, set to the observed proportion of times that the subjects chose option A. However, such an approach can overfit the empirical choices. We therefore instead assumed that the observed data were drawn from a probabilistic data-generating process, and produced a probability distribution over $p_{\text{subject}}(A)$. To do this, we set an uninformative prior over $p_{\text{subject}}(A)$, and treated the past NHP choice data for each experiment as observations generated from a binomial distribution (since there were two outcomes in the form of the two food rewards the subject could chose) with an unknown $p_{\text{subject}}(A)$.

Conditioning on $p_{\text{subject}}(A)$ (based on the observed choice data reported in these experiments), we performed Bayesian inference over $p_{\text{model}}(A)$ and r (with a uniform prior over both), and marginalized over all possible values of $p_{\text{model}}(A)$ to result in a posterior distribution over r (visualized in Fig. 6b).

There are a couple of approaches to obtain a point estimate of r. One possibility would be to use the maximum a-posteriori (MAP) estimate for $\hat{r}$ (i.e., the highest point in Fig. 6b). However, MAP estimates under a uniform prior are equivalent to a maximum likelihood estimate, which is prone to overfitting a distribution, particularly when little data is available. Here, we instead opted to take the expected value of the posterior distribution, as this estimator minimizes the expected error of the point estimate (65)—a preferred approach in situations like this where data are limited.

5.4.2. Fitting reliance to data

Our final analysis tested how different reliance parameters affected model performance. To achieve this, we calculated the probability of the data for each model, testing reliance values between 0 and 1 by intervals of 0.01. These are the results shown in Fig. 5. The *Egocentric* models are insensitive to the reliance parameter because, by construction, they do not have any social representations. The *Behavioral* models (*Food-directed behavior* and *Location-directed behavior*) and the *True belief* model showed a similar structure to one another: the stronger the assumption that primates always use these representations, the better these models can explain the empirical data. The other two models in the *Mentalistic* family (*Uncertain representation* and *Partial representation*) were best able to explain the empirical data under the assumption that subjects rely on their social representations more than half of the time, but critically, not all of the time. Finally, the *Ignorance representation* model's ability to explain the empirical data rapidly decreased when it was posited that subjects always relied on their underlying social representations, because the model expected the experiments to show a much stronger signal than they did.

6. Materials Availability Statement

All model and analysis code is available at: https://osf.io/qjn9m/?view_only=a8ede1cffb684452af9b2c54cb5ac15c.

7. Acknowledgments

We thank Mel Andrews, Amanda Royka, and Max Siegel for helpful comments. This work was supported by NSF CAREER award BCS-2045778 awarded to JJE and NSF SBE Postdoctoral Research Fellowship 2104589 awarded to DJH.

References

[1] Mitani JC, Watts DP. Why do chimpanzees hunt and share meat? Animal Behaviour 2001;61(5):915–24.

[2] Mitani JC. Cooperation and competition in chimpanzees: current understanding and future challenges. Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews 2009;18(5):215–27.

[3] Muller MN, Mitani JC. Conflict and cooperation in wild chimpanzees. Advances in the Study of Behavior 2005;35:275–331.

[4] Perry S. Social traditions and social learning in capuchin monkeys (cebus). Philosophical Transactions of the Royal Society B: Biological Sciences 2011;366(1567):988–96.

[5] Cheney DL, Seyfarth RM. Baboon metaphysics. In: Baboon Metaphysics. University of Chicago Press; 2008,.

[6] Brent LJ, Chang SW, Gariépy JF, Platt ML. The neuroethology of friendship. Annals of the New York Academy of Sciences 2014;1316(1):1–17.

[7] Cheney DL, Seyfarth RM. How monkeys see the world: Inside the mind of another species. University of Chicago Press; 2018.

[8] Tomasello M, Carpenter M, Call J, Behne T, Moll H. Understanding and sharing intentions: The origins of cultural cognition. Behavioral and brain sciences 2005;28(5):675–91.

[9] Gweon H. Inferential social learning: Cognitive foundations of human social learning and teaching. Trends in Cognitive Sciences 2021;25(10):896–910.

[10] Kiley Hamlin J, Ullman T, Tenenbaum J, Goodman N, Baker C. The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. Developmental Science 2013;16(2):209–26.

[11] Jara-Ettinger J, Rubio-Fernandez P. Quantitative mental state attributions in language understanding. Science advances 2021;7(47):eabj0970.

[12] Onishi KH, Baillargeon R. Do 15-month-old infants understand false beliefs? science 2005;308(5719):255–8.

[13] Liu S, Ullman TD, Tenenbaum JB, Spelke ES. Ten-month-old infants infer the value of goals from the costs of actions. Science 2017;358(6366):1038–41.

[14] Kampis D, Southgate V. Altercentric cognition: How others influence our cognitive processing. Trends in Cognitive Sciences 2020;24(11):945–59.

[15] Rubio-Fernández P, Mollica F, Ali MO, Gibson E. How do you know that? automatic belief inferences in passing conversation. Cognition 2019;193:104011.

[16] Rosati AG, Santos LR, Hare B. Primate social cognition: Thirty years after premack and woodruff. In: Primate neuroethology. Oxford University Press; 2010, p. 117–43.

[17] Krupenye C, Call J. Theory of mind in animals: Current and future directions. Wiley Interdisciplinary Reviews: Cognitive Science 2019;10(6):e1503.

[18] Call J, Tomasello M. Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences 2008;12(5):187–92. doi:https://doi.org/10.1016/j.tics.2008.02.010.

[19] Penn DC, Povinelli DJ. On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. Philosophical Transactions of the Royal Society B: Biological Sciences 2007;362(1480):731–44.

[20] Andrews K. Do apes read minds?: Toward a new folk psychology. mit Press; 2012.

[21] Andrews K. Apes track false beliefs but might not understand them. Learning & Behavior 2018;46:3–4.

[22] Heyes CM. Theory of mind in nonhuman primates. Behavioral and brain sciences 1998;21(1):101–14.

[23] Baker CL, Jara-Ettinger J, Saxe R, Tenenbaum JB. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour 2017;1:0064.

[24] Gerstenberg T, Goodman ND, Lagnado DA, Tenenbaum JB. A counterfactual simulation model of causal judgments for physical events. Psychological review 2021;128(5):936.

[25] Ho MK, Saxe R, Cushman F. Planning with theory of mind. Trends in Cognitive Sciences 2022;26:959–71.

[26] Phillips J, Buckwalter W, Cushman F, Friedman O, Martin A, Turri J, et al. Knowledge before Belief. Behavioral and Brain Sciences 2020;44:e140.

[27] Luo Y, Johnson SC. Recognizing the role of perception in action at 6 months. Developmental Science 2009;12(1):142–9.

[28] Penn DC, Povinelli DJ. On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. Philosophical transactions of the Royal Society of London Series B, Biological sciences 2007;362(1480):731–44.

[29] Povienlli DJ, Vonk J. Chimpanzee minds: suspiciously human? Trends in Cognitive Sciences 2003;7(4):157–60.

[30] Lurz RW, Kanet S, Krachun C. Animal mindreading: A defense of optimistic agnosticism. Mind & Language 2014;29(4):428–54.

[31] Santos LR, Nissen AG, Ferrugia JA. Rhesus monkeys, macaca mulatta, know what others can and cannot hear. Animal Behaviour 2006;71(5):1175–81.

[32] Melis AP, Call J, Tomasello M. Chimpanzees (Pan troglodytes) conceal visual and auditory information from others. Journal of Comparative Psychology 2006;120(2):154–62.

[33] Noë R, de Waal FB, van Hooff JA. Types of dominance in a chimpanzee colony. Folia Primatologica 1980;34(1-2):90–110.

[34] Hare B, Call J, Tomasello M. Chimpanzees know what conspecifics do and do not see. Animal Behaviour 2000;59:771–85.

[35] Bräuer J, Call J, Tomasello M. Chimpanzees really know what others can see in a competitive situation. Animal Cognition 2007;10(4):439–48.

[36] Canteloup C, Piraux E, Poulin N. Do tonkean macaques (macaca tonkeana) perceive what conspecifics do and do not see? PeerJ 2016;:1–21.

[37] Hare B, Call J, Tomasello M. Chimpanzees deceive a human competitor by hiding. Cognition 2006;101(3):495–514.

[38] Martin A, Santos LR. What cognitive representations support primate theory of mind? Trends in cognitive sciences 2016;20(5):375–82.

[39] Hutto DD. Folk psychological narratives: The sociocultural basis of understanding reasons. MIT press; 2012.

[40] Gordon RM. The simulation theory: Objections and misconceptions. Mind & language 1992;7:11–34.

[41] Jara-Ettinger J. Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences 2019;29:105–10.

[42] Samson D, Apperly IA, Braithwaite JJ, Andrews BJ, Bodley Scott SE. Seeing it their way: evidence for rapid and involuntary computation of what other people see. Journal of Experimental Psychology: Human Perception and Performance 2010;36(5):1255.

[43] Tomasello M, Call J, Hare B. Five primate species follow the visual gaze of conspecifics. Animal Behaviour 1998;55(4):1063–9.

[44] MacLean EL, Hare B. Bonobos and chimpanzees infer the target of another's attention. Animal Behaviour 2012;83(2):345–53.

[45] Rosati AG, Arre AM, Platt ML, Santos LR. Rhesus monkeys show human-like changes in gaze following across the lifespan. Proceedings of the Royal Society B: Biological Sciences 2016;283(1830):20160376.

[46] Drayton LA, Santos LR. Rhesus monkeys show human-like changes in gaze following across the lifespan. Animal Behaviour 2017;147:193–9.

[47] Bettle R, Rosati AG. Flexible gaze-following in rhesus monkeys. Animal Cognition 2019;22(5):673–86.

[48] Hare B. Can competitive paradigms increase the validity of experiments on primate social cognition? Animal Cognition 2001;4(3-4):269–80.

[49] Drayton LA, Santos LR. What do monkeys know about others' knowledge? Cognition 2018;170:201–8.

[50] Horschler DJ, Santos LR, MacLean EL. Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. Cognition 2019;190:72–80.

[51] Horschler DJ, Santos LR, MacLean EL. How do non-human primates represent others' awareness of where objects are hidden? Cognition 2021;212:104658.

[52] Kaminski J, Call J, Tomasello M. Chimpanzees know what others know, but not what they believe. Cognition 2008;109(2):224–34.

[53] Krupenye C, Kano F, Hirata S, Call J, Tomasello M. Great apes anticipate that other individuals will act according to false beliefs. Science 2016;354(6308):110–4.

[54] Kano F, Call J, Krupenye C. Primates pass dynamically social anticipatory-looking false-belief tests. Trends in Cognitive Sciences 2020;24(10):777–8.

[55] Horschler DJ, MacLean EL, Santos LR. Do non-human primates really represent others' beliefs? Trends in Cognitive Sciences 2020;24(8):594–605.

[56] Horschler DJ, MacLean EL, Santos LR. Advancing gaze-based research on primate theory of mind. Trends in Cognitive Sciences 2020;24(10):778–9.

[57] Povinelli DJ, Eddy TJ, Hobson RP, Tomasello M. What Young Chimpanzees Know about Seeing. Monographs of the Society for Research in Child Development 1996;61(3).

[58] Bellman R. A Markovian Decision Process. Journal of Mathematics and Mechanics 1957;6(5):679–84.

[59] Jara-Ettinger J, Schulz L, Tenenbaum J. The Naive Utility Calculus as a unified, quantitative framework for action understanding. Cognitive Psychology 2020;123:101334.

[60] Baker CL, Saxe R, Tenenbaum JB. Action understanding as inverse planning. Cognition 2009;113(3):329–49.

[61] Gergely G, Csibra G. Teleological reasoning in infancy: The naïve theory of rational action. Trends in Cognitive Sciences 2003;7(7):287–92.

[62] Rochat MJ, Serra E, Fadiga L, Gallese V. The evolution of social cognition: Goal familiarity shapes monkeys' action understanding. Current Biology 2008;18(3):227–32.

[63] Jern A, Lucas CG, Kemp C. People learn other people's preferences through inverse decision-making. Cognition 2017;168:46–64.

[64] Lucas CG, Griffiths TL, Xu F, Fawcett C, Gopnik A, Kushnir T, et al. The child as econometrician: A rational model of preference understanding in children. PloS one 2014;9(3):e92160.

[65] Jaynes ET. Probability theory: The logic of science. Cambridge university press; 2003.