Cognition 278 (2027) 106698

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

What primates know about other minds and how often they use it: A computational approach to comparative theory of mind

Marlene D. Berke [a,b,*,1], Daniel J. Horschler [a,1], Amanda Royka [a], Laurie R. Santos [a], Julian Jara-Ettinger [a]

[a] Yale University, United States of America
[b] Massachusetts Institute of Technology, United States of America

ARTICLE INFO

Dataset link: https://doi.org/10.17605/OSF.IO/QJN9M

Keywords:
Comparative cognition
Computational modeling
Social cognition
Theory of mind

ABSTRACT

Can non-human primates (NHPs) represent other minds? Answering this question is difficult because primates can fail tasks due to a lack of motivation or succeed through simpler strategies. Here, we address these challenges through a computational theory-testing framework for NHP Theory of Mind. In this framework, each theory combines a proposed social representation with a parameter for how often it is used. Fitting these models makes explicit an assumption that existing accounts leave unstated: how often NHPs must rely on a posited representation for that account to fit the data. This allows us to move beyond dichotomous positions about the presence or absence of Theory of Mind and instead analyze graded patterns of behavior as a combination of cognitive representations and their use. We apply this approach to one of the most foundational and well-studied aspects of Theory of Mind: the relation between seeing and knowing. Our results show that only theories in which NHPs have some representation of other minds can capture the qualitative pattern of successes and failures across five classic perspective-taking paradigms. However, these theories vary in their reliance on their representations, each showing significantly lower reliance than a human baseline. These results suggest that human and NHP social cognition differ in terms of reliance and possibly also in terms of representational complexity.

Like humans, non-human primates (NHPs) have rich social behaviors. They live in complex social groups, empathize with and help each other, and develop lifelong relationships (Cheney & Seyfarth, 2008; Lewis et al., 2023; Muller & Mitani, 2005; Warneken et al., 2007). Yet, their social life is undeniably simpler than ours (Hall & Brosnan, 2017; Hoppitt et al., 2008; Musgrave et al., 2016; Tomasello et al., 2005). NHPs share knowledge only in limited ways: they do not intentionally deceive one another, and they do not build cultures of human-level complexity. What cognitive differences explain the unique complexity of human social life?

A common hypothesis focuses on *Theory of Mind*—the ability to interpret and predict others' behavior by representing unobservable mental states like beliefs and desires. In humans, Theory of Mind supports cooperation, communication, and moral reasoning (Gweon, 2021; Rubio-Fernández et al., 2025; Young et al., 2007). Its foundational components emerge in infancy (Csibra et al., 2003; Gergely & Csibra, 2003; Liu et al., 2017; Repacholi & Gopnik, 1997), and many important mental-state inferences become automatic in adults (Kampis & Southgate, 2020; Rubio-Fernández et al., 2019). Differences between human and NHP Theory of Mind (hereafter NHP ToM) are therefore a natural candidate to explain the relative simplicity of NHP social life.

This hypothesis has led to extensive research on comparative Theory of Mind (see Call & Tomasello, 2008; Krupenye & Call, 2019; Rosati et al., 2010, for reviews), but there is little consensus on what NHPs understand about others' mental states (compare, e.g., Krupenye et al., 2016; Martin & Santos, 2016). This stems from two fundamental challenges in comparative research.

The first challenge is practical. When NHPs fail a ToM task, does the failure reflect the absence of a capacity, or a failure to *rely* on it? NHPs might be able to represent other minds, but not do so in a given task. This could be due to cognitive load, task complexity, or lack of motivation. For example, some research has argued that NHPs have a ToM, but only use it selectively in adversarial situations with "social drama" (Hare & Tomasello, 2004; Kano et al., 2020). This would explain why only some tasks elicit ToM-like behavior and why effect sizes are weaker than those seen in humans.

* Corresponding author at: Massachusetts Institute of Technology, United States of America.
E-mail address: berkem@mit.edu (M.D. Berke).
[1] These authors made equal contributions and are listed in alphabetical order.

https://doi.org/10.1016/j.cognition.2026.106698
Received 10 April 2026; Received in revised form 25 August 2026; Accepted 27 August 2026
Available online 12 September 2026
0010-0277/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

The second challenge is theoretical. Success on ToM tasks can often be explained by multiple accounts that posit different representations while predicting the same qualitative outcomes. This is a common challenge in comparative research: for nearly every study on Theory of Mind, there are competing accounts that explain the qualitative data equally well (Andrews, 2012, 2018; Heyes, 1998; Penn & Povinelli, 2007).

Our work seeks to address both challenges through a computational approach. We implement a series of models, each positing a different representation that NHPs might use in social tasks. Because each account is formalized computationally, it makes not just qualitative predictions about whether NHPs should succeed or fail, but precise, quantitative predictions about how large the effect should be. This allows us to compare competing accounts that make the same qualitative predictions against each other and evaluate whether their predicted effect sizes match the data.

This approach has been successful in the study of human social cognition, helping distinguish even between subtle variations of the same theory (e.g., Jara-Ettinger et al., 2020). However, this approach cannot be directly applied to NHPs. A model's predictions might misalign with primate data not because the posited representations are wrong, but because NHPs were not relying on their social representations in that task. We therefore introduce the construct of *reliance* to capture the frequency with which NHPs use the representation that a theory ascribes to them. Our notion of reliance is intentionally broad, aggregating over several factors that might cause NHPs to under-use their capacities, including motivation, cognitive load, and mood. We return to this point in the General Discussion.

This approach lends two key advantages to our framework. First, we can falsify theories whose model formalizations outright fail to capture NHP behavior in qualitative ways. Second, for theories that make similar qualitative predictions, we can compare predicted effect sizes to NHP data. If NHPs perform better than a model predicts, this suggests their behavior is more complex than this set of representations can support. In that case, the account can be discarded. By contrast, when NHPs perform worse than a model predicts, we can estimate how often NHPs would need to rely on the posited representation for the account to hold. In doing so, our framework makes explicit assumptions about reliance that would otherwise remain hidden. This may reveal that some accounts are only plausible under a specific and surprisingly low degree of reliance. While our approach describes a general framework that can be applied to many different types of representations, we begin by describing the specific domain that this work explores.

0.1. Scope and paradigms

Theory of Mind is a complex cognitive system with many subcomponents (including mechanisms for inference, prediction, explanation, and even planning over other minds; Ho et al., 2022; Wang et al., 2025). Here we focus on one well-studied component: the ability to determine what others see and know based on their visual perspective (see Phillips et al., 2021, for review). This capacity to link seeing and knowing is one of the most basic components of ToM, emerging early in human infancy (Luo & Johnson, 2009), and whether NHPs share it has been much debated (Call & Tomasello, 2008; Halina, 2017; Lurz et al., 2014; Rosati et al., 2010).

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 in 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, they can use this to strategically decide which apple to pursue. Tasks like this were used in now-classical work to show that chimpanzees (*Pan troglodytes*) will preferentially approach food rewards that are hidden from conspecifics, providing some of the first evidence for visual perspective-taking in NHPs (Hare et al., 2000). This work has since been extended to show more complex forms of perspective-taking, including evidence that NHPs know what others can hear (Melis et al., 2006; Santos et al., 2006).

To explore possible representations, we developed seven computational models of varying cognitive complexity (Fig. 1b–h; Table 2), explained in detail in the following section. Each model is instantiated as an agent that can complete basic primate experiments in two-dimensional grid worlds. We then fit each model to the empirical record of NHP visual perspective-taking, incorporating the reliance parameter to account for gaps between predicted and observed behavior.

Our framework makes two key contributions to the study of comparative Theory of Mind. First, and most centrally, it makes explicit an assumption that existing accounts of NHP social cognition leave unstated: how often NHPs must rely on the representations a given account posits in order for that account to explain the observed data. Rather than proposing a new theory of NHP social cognition, we take existing accounts and implement them as computational models to measure the reliance commitments they implicitly carry. This reveals that some accounts are only plausible under specific, and sometimes surprisingly low, degrees of reliance. Second, because the approach is not tied to visual perspective-taking, it can be applied to other debates within comparative cognition, including other aspects of ToM and beyond (e.g., NHP intuitive physics), laying the groundwork for a broader use of computational methods in comparative cognitive science.

1. Overview of experimental paradigms and modeling

1.1. Experimental paradigms

Our work focused on 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 (Noë et al., 1980), these paradigms create 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 also see a second food reward that is hidden from the competitor's view by a barrier. The Open-Hidden paradigm is similar except that the two rewards are equidistant from the subject. In both of these paradigms, NHP subjects 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 path that is hidden from or visible to the competitor. Because reaching along the hidden path enables the subject to escape detection, it is the better choice, and NHPs are indeed more likely to choose it.

The final two paradigms are controls. In the Hidden-Hidden paradigm, both food rewards are visible to the subject but hidden from the competitor. The Open-Transparent paradigm is identical to the Open-Hidden paradigm except that the barrier is transparent (so the subject and competitor see both food rewards). In both of these paradigms, NHPs show no systematic preference for either food reward.

1.2. Computational framework

To test what social representations underlie NHP behavior in these tasks, we implemented seven computational models that vary in representational complexity. To validate our approach, we first implemented three non-mentalistic models, which assume that NHPs predict others'

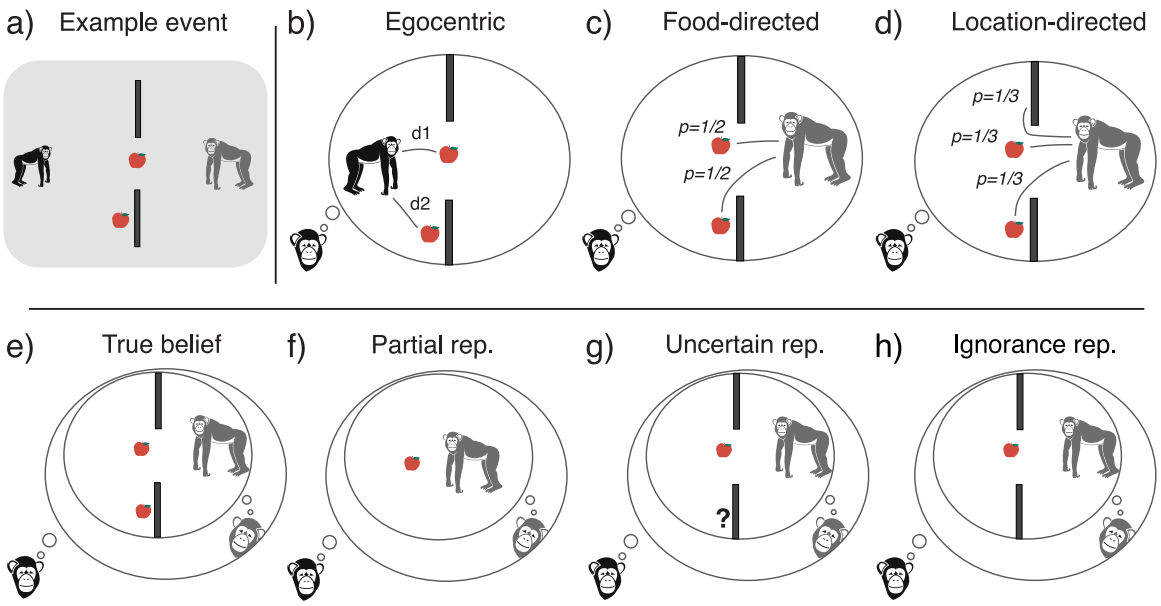

Fig. 1. (a) A hypothetical competitive task used to test the relation between seeing and knowing. Here, the subject (small chimpanzee, black) is in front of a dominant conspecific (big chimpanzee, gray) in a room with two apples and two opaque barriers, such that both chimpanzees see the apple between them, but only the subject sees the apple in front of the barrier. The dominant chimpanzee will take all the food that they know about, so the subject's only hope to obtain food is to approach the apple that is hidden from the competitor. If the subject approaches the same apple as the competitor, they will not only lose the contest for the apple, but they risk physical aggression. (b–h) Visual illustration of our seven computational models, ordered by complexity. The models fall into three broad families: non-mentalistic (b–d), semi-mentalistic (e–g), and full ToM (h). Throughout, *p* refers to probabilities, and *d* refers to distances.

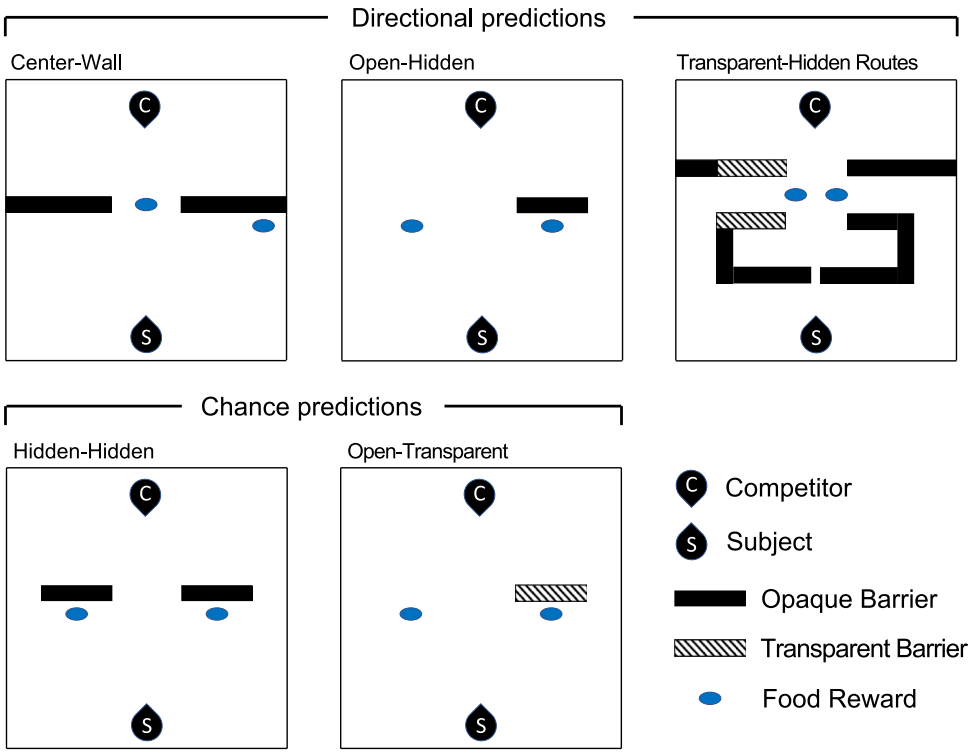

Fig. 2. Depictions of the experimental setup for each of the five paradigms we used to compare our computational models to NHP behavior. The top row shows paradigms with directional predictions: if subordinate NHPs represent others' knowledge, then they should preferentially approach the reward (or take the route) hidden from the competitor. The bottom row shows control paradigms. When the two rewards are simultaneously known or hidden from the competitor, the subject should show no reliable preference.

Table 1
Papers and experiments used to evaluate our computational models. Hare et al. (2000) Exp 3 and 4 had multiple conditions that used more than one paradigm and are therefore repeated. All experiments tested chimpanzees (*Pan troglodytes*) except for Canteloup et al. (2016), which tested Tonkean macaques (*Macaca tonkeana*).

Paradigm	Paper	Exp
Center-Wall	Hare et al. (2000)	E1
Open-Hidden	Bräuer et al. (2007)	E2
	Canteloup et al. (2016)	E1
	Hare et al. (2000)	E2–E4
Transparent-Hidden Routes	Hare et al. (2006)	E2
	Melis et al. (2006)	E1
Hidden-Hidden	Hare et al. (2000)	E3–E4
Open-Transparent	Hare et al. (2000)	E5

behavior based on simple rules. These were inspired by typical baseline alternative hypotheses that are considered null hypotheses. Given the success of NHPs in the tasks considered here, these models should fail to capture the data and under-estimate NHP performance. They also serve to ensure that not just any theory can be made plausible by manipulating the notion of reliance. Therefore, these models serve as baselines for validating our framework.

The remaining four models capture existing theories of NHP ToM-like representations. One model implements a full ToM, and the other three are best described as "semi-mentalistic", proposing that NHPs track others' perspectives and attribute mental states, but those representations do not rise to the level of decoupled, non-factive mental states such as ignorance and false beliefs. Thus, the theories we present are a representative sample from the space of existing theories, varying across levels of complexity. All model and analysis code are available in our OSF repository: https://doi.org/10.17605/OSF.IO/QJN9M.

Given a set of possible rewards that the NHP can pursue (see Fig. 2 for paradigms), each model aims to predict the probability that an NHP will choose to approach reward R, given by

$$p_{\text{Predictive}}(R) = 1 - p_{\text{Competitor}}(R) \tag{1}$$

where $p_{\text{Competitor}}(R)$ is the probability that the competitor will also pursue reward R according to the subject's cognitive representations. This equation captures the idea that the subject wants to avoid the reward that the dominant conspecific will choose. The subject will choose reward R with probability 1 only when certain that the competitor will not go for it (i.e., $p_{\text{Competitor}}(R) = 0$). Conversely, the subject will never choose reward R if confident that the competitor will go for it (i.e., $p_{\text{Competitor}}(R) = 1$). This graded response rule is well-motivated: animal decision-making often shows probability-tracking choice patterns (e.g., Herrnstein, 1961; but c.f. Saldana et al., 2022), it follows standard practice in the inverse-planning tradition (e.g., Baker et al., 2017, 2009; Jara-Ettinger et al., 2020), and it avoids discontinuities in model behavior. This formulation is general to all of the models, with each model specifying a different mechanism through which the NHP predicts the competitor's choice $p_{\text{Competitor}}$.

In some cases, NHPs might not compute or rely on $p_{\text{Competitor}}(R)$ at all. In those cases, we assume that the NHP acts egocentrically, as they would in the absence of a competitor, denoted by $p_{\text{Egocentric}}(R)$. The propensity to predict the competitor's behavior is determined by *reliance*, which we denote as r. In our formulation, the NHP's proclivity to choose reward R is given by

$$p_{\text{Subject}}(R) = r p_{\text{Predictive}}(R) + (1 - r) p_{\text{Egocentric}}(R). \tag{2}$$

According to this equation, the NHP acts based on its prediction of the competitor's behavior (specified in Eq. (1)) with probability r (reliance). Otherwise (with probability $1-r$), the NHP acts egocentrically, without integrating predictions about the competitor. Eq. (2) specifies a propensity for each option available to the subject; the probability of choosing a given option is obtained by normalizing these propensities across the options. See SM Section 1.1 for details.

Thus, in this formulation, reliance r is a free parameter that can be fit to NHP data, but its value depends on the posited mechanism for predicting competitor behavior. Thus, each theory of primate social representations produces a different $p_{\text{Predictive}}(R)$ estimate (by predicting different competitor behavior), and we then adjust the reliance to maximize the model's alignment with the empirical data.

In the remainder of this section, we first discuss how we compute egocentric behavior, $p_{\text{Egocentric}}(R)$, and then go through the models, each specifying a different mechanism for predicting the competitor's behavior $p_{\text{Competitor}}(R)$.

1.2.1. Modeling egocentric behavior

To compute $p_{\text{Egocentric}}(R)$, we implemented the experimental paradigms as Markov Decision Processes (MDP; Bellman, 1957). MDPs are a general framework for modeling how agents take sequential actions in an environment to obtain rewards. Because MDPs are a popular framework in cognitive science and social cognition, here we assume some basic knowledge of this framework. A complete introduction to MDPs can be found in Sigaud and Buffet (2013), and a tutorial for MDPs applied to social reasoning can be found in Jara-Ettinger et al. (2024). For the simple paradigms tested in this work, this MDP formulation is more machinery than strictly necessary: for example, $p_{\text{Egocentric}}(R)$ could be calculated through simpler, hand-specified rules (e.g., through a softmax over distance). We nonetheless route these quantities through explicit MDP simulations because doing so keeps the framework general rather than bespoke to these five paradigms; the same machinery extends, without modification, to gridworlds with more complex geometry, multi-step or path-dependent trajectories, and full physical-planning problems, thus laying the groundwork for future work modeling more complex tasks.

Using MDPs, we represent each experimental paradigm as a simple grid world where the NHP can sequentially move in any of the four cardinal directions (up, down, left, right) as well as diagonally (up-left, up-right, down-left, down-right). Each movement incurs a small cost to induce an expectation of efficient action (by making longer action paths more costly), and reaching a food produces a reward.

The cost of action was sampled from an exponential distribution with $\lambda = 100$ (mean cost = 0.01) to ensure that results did not depend on any particular value of movement cost. This value determined the cost of moving in one of the four cardinal directions, with the cost of moving in any diagonal direction given by $\sqrt{2}c$ (by the Pythagorean theorem).

The reward was set to be large ($R = 50$) so that it would always outweigh movement costs (i.e., to avoid situations where the food is not rewarding enough to justify the cost of getting it; as this has never been observed in the experimental paradigms under consideration). Note that exact reward values have important consequences in environments where different objects have different rewards, but this is not the case in our paradigms, where all food rewards were identical.

MDPs further require that rewards be temporally discounted, such that an immediate reward is more valuable than a reward obtained

Table 2
Summary of computational models. See SM Section 1 for technical presentation.

Family	Model	Summary
Non-mentalistic	Egocentric	NHP always acts egocentrically. No social prediction mechanisms.
	Food-directed	Competitor has an equal probability of going to any food in the room.
	Location-directed	Competitor has an equal probability of going to any actual or possible food location
Semi-mentalistic	True belief	Assume shared knowledge.
	Partial representation	Represent only what is in shared view.
	Uncertainty representation	Competitor knows about everything they see and may or may not know about any object they do not see.
Full ToM	Ignorance representation	Accurate representation of what the competitor knows and does not know.

in the future. The temporal discount factor was set to slightly favor immediate rewards over future ones ($\gamma = 0.95$).

Equipped with this formulation, MDPs make it possible to calculate which actions an agent should take to obtain the highest possible reward (called a policy). Standard algorithms focus on obtaining the optimal solution, no matter the difficulty of the problem. In reality, as the values of two alternative food rewards become more similar, it should be harder to reliably identify which one is better (e.g., it is obvious that a reward one step away is better than an identical reward fifty steps away, but the choice would be far harder if the two rewards were, say, thirty-three and thirty-four steps away in different directions). To account for this, we made the MDP policy probabilistic, using the standard approach of softmaxing the value functions (Baker et al., 2017; Jara-Ettinger et al., 2020; Jern et al., 2017; Lucas et al., 2014). We use a small but non-zero temperature parameter of $\tau = 0.05$ so that model fits gradually degrade with the amount of conflicting evidence, rather than a single inconsistent choice driving the likelihood to zero. At this setting, in the Center-Wall paradigm (Fig. 2), the probability of the subject pursuing the center food (value 50, three steps away) vs. the food on their side of the barrier (value 50, four steps away) is about 0.51. This is in keeping with the spirit of the experiment, where small differences in physical distances are not considered main drivers of behavior (Hare et al., 2000); instead, behavior depends on the NHPs' expectations of the relationship between seeing and knowing. We fix $\tau = 0.05$ throughout and later examine the results of varying this value (Fig. 5).

1.3. Modeling competitor behavior

So far, our framework has specified (1) the general social behavior where the NHP avoids the reward they expect the dominant competitor will choose, (2) a tradeoff between egocentric and social behavior (governed through a reliance parameter), and (3) how egocentric behavior is determined. The last remaining component is how the NHP makes predictions about their competitor. We explored this through seven computational models, summarized in Table 2. While these models can be thought of as points in a continuum of representational complexity, they can be roughly organized into three broad categories of accounts: *Non-mentalistic*, *Semi-mentalistic*, and *Full Theory of Mind*. For simplicity, here we introduce the high-level intuition behind each model, and a formal presentation is available in SM Section 1.

1.3.1. Non-mentalistic models

The first family of accounts constitutes null hypotheses: they are rule-like ways in which NHPs could approach these tasks without representing other minds. We started with the simplest account, the *Egocentric* model, where NHPs entirely ignore the competitor. According to this model, NHPs simply pursue the food reward reachable by the shortest path (Fig. 1b), determined using the egocentric behavior model described in the general framework.

We next considered two accounts wherein NHPs predict behavior by relying on simple behavioral cues. These models are more sophisticated in that they require the subject to estimate how their competitor might behave. Thus, these models (and all models in more complex families) also rely on MDPs to model how the competitor will move in space. While MDPs are often used to model first-person behavior, 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 (Baker et al., 2017, 2009; Jara-Ettinger et al., 2020). Critically, the MDP framework allows agents to be modeled as efficient actors by approximately following the MDP's optimal policy (where the softmaxing τ parameter introduced above also modulates the strength of this expectation), thus implementing the assumption that agents move efficiently toward their goals—an assumption that both human and non-human primates share (Gergely & Csibra, 2003; Rochat et al., 2008).

In the *Food-directed behavior* model, the subject expects the competitor to pursue one of the food rewards (possibly learned from experience seeing conspecifics go toward food) but lacks a mechanism for predicting which one, therefore placing an equal probability over each food reward (Fig. 1c), regardless of whether they are visible or not. In these paradigms, because there are always two food rewards, this model chooses each with 50% probability. This model can therefore be considered as a baseline representing chance performance.

The *Location-directed behavior* model extends behavioral expectations to include the expectation that competitors might also check locations that they cannot see (which could also be learned from experience observing conspecifics without necessarily representing epistemic states). This model therefore places equal probability over all food rewards and places where food could be located (even if no food is there; Fig. 1d).

1.3.2. Semi-mentalistic models

The *semi-mentalistic* family consists of three models that represent the competitor's mind, but lack full human-like mental state representations. The *True belief* model is a classical model of how agents might be able to reason about others in an over-simplified but useful way,

by attributing their own knowledge to others (Horschler et al., 2020b; Wellman & Woolley, 1990; Wimmer & Perner, 1983). In the context of our task, this corresponds to NHPs assuming that their competitor shares their knowledge (Fig. 1e). However, even if 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 the food closest to them, which could be different for the subject and the competitor).

The *Partial representation* model implements a version of the awareness relations account of NHP ToM (Martin & Santos, 2016). In this account, primates have a first-person representation of the world but cannot instantiate parallel representations for others (leaving them unable to represent mental states that differ from their own). Unlike the *True belief* model, however, primates do not simply assume others know everything they know. Instead, they attribute awareness only of objects in the shared visual field, and anything outside it is not represented in the other mind at all. This creates a simplified social representation where the subject represents only the shared visual field and predicts the competitor's behavior within it (Fig. 1f). Therefore, objects and paths hidden from the competitor are absent from this representation. While this account has been used to explain NHP failures on false belief tasks (Martin & Santos, 2016), this is the first test against the visual perspective-taking literature.

In the *Uncertainty representation* model, the NHP assumes the competitor is aware of any object in their visual field, but is uncertain whether the competitor is also aware of objects outside their field of view. This model differs from the *Partial representation* model in that the subject considers the possibility that the conspecific might know about the hidden reward (but is unsure whether or not they know about it). This model therefore integrates the two epistemic hypotheses about the competitor (knows or does not know about the hidden object, using a uniform prior; Fig. 1g) to build predictions for every possible location that is not in their visual field. This model is based on recent discussions about what representations of "not knowing" might look like in NHPs (Phillips et al., 2021; Royka & Jara-Ettinger, 2021; Royka & Santos, 2025; Townrow & Krupenye, 2025). This uncertainty account captures one way in which NHPs might track and make predictions based on unshared visual access without having fully decoupled ignorance representations.

Full Theory of Mind model

Finally, the *Full ToM* family consists of theories of human-like Theory of Mind. Although there are a variety of different proposed theories of what full human-like ToM consists of (e.g., Gordon, 1992; Hutto, 2012; Jara-Ettinger, 2019), these theories only make different predictions in complex cases that go beyond those captured in the paradigms considered here. So for this family, we include only an *Ignorance representation* model. This model implements a decoupled representation of the competitor's visual field: the subject uses one grid world to represent the beliefs of their competitor and predict their actions and a separate grid world for planning their own actions. Thus, the model's architecture separates the subject's and the competitor's knowledge and is, in principle, capable of representing not just ignorance but false beliefs and other non-factive mental states. In this way, the model is more general than the paradigms on which it is tested here, which only test ignorance and knowledge.[2]

2. Results

Our analysis consisted of two stages. We first examined how each model performs under full reliance to compare predicted to empirical effect sizes. Then, we added the reliance parameter and fit it to empirical data to reveal the relation between posited representations and their reliance.

[2] This also means that, even if the *Ignorance representation* model fits well on these paradigms, NHPs might still behave quite differently on false belief paradigms.

2.1. Model performance under full reliance

We analyzed the exact effect sizes that each model predicts to reveal whether a model, even when it makes the correct directional prediction, may lack explanatory power by either over- or under-predicting an effect size relative to observed NHP behavior. Therefore, we directly compared the predicted and empirical effect sizes for each paradigm.

To make these effects easier to interpret, we ran a simple online experiment testing human participants on analogous visual perspective-taking tasks: choosing which candy to take in the presence of a bully. The full experimental procedure for the human experiment is available in SM Section 2. This experiment grounds our subsequent comparison between humans and NHPs on empirical data and effect sizes.

Fig. 3 shows each model's performance across the five paradigms, overlaid with human and NHP results. In each plot, black lines represent model predictions and shaded regions represent empirical data, such that a closer fit between the black line and the shaded region indicates better model-data alignment.

The three mentalistic models differ in the strength of the effects they predict because of how each represents the competitor in the Center-Wall, Open-Hidden, and Transparent-Hidden Routes paradigms. *Ignorance rep.* confidently represents the competitor as unaware of anything hidden from its view, predicting that the competitor will neither pursue hidden rewards nor detect the subject along hidden routes; this yields the strongest, most consistent effects. *Partial rep.* reaches nearly the same predictions in most paradigms by simply omitting hidden objects and regions from its representation of the competitor's mind, but in the Transparent-Hidden Routes paradigm the hidden route is absent from the representation, so the model makes no prediction at all about whether the competitor will detect the subject there, yielding a slightly weaker effect. *Uncertainty rep.* instead explicitly represents the competitor as uncertain, averaging over the possibilities that the competitor does and does not know about the hidden reward or route; this averaging systematically yields the most moderate, least extreme predictions. A detailed walk-through of each model's predictions is available in SM Section 3.

For human data (top row, pink), all non-mentalistic models under-predicted effect sizes (i.e., the pink region exceeds the black contour), and so did the *True belief* model in the semi-mentalistic family. In contrast, the other two semi-mentalistic models and the full ToM model all showed a closer alignment with human behavior.

For NHP data (bottom row, green), a different pattern emerged. Here, the four simplest models (the non-mentalistic models and the *True belief* model) under-predicted the observed effects, while the three most complex models over-predicted the effects in some paradigms.

To confirm this visual assessment quantitatively, we would ideally calculate each model's explanatory power over the empirical data, $P(D|M)$. However, the trial-level performance data needed for this analysis is not available for every paradigm (see SM Section 4.1 for details), so we begin by computing the root mean squared error (RMSE) between model predictions and empirical data, allowing us to include all paradigms. This analysis mimics the visual assessment by weighting each paradigm equally, regardless of how much data is available per paradigm. Additional information and full results are available in SM Section 3.2.1.

The RMSE analysis confirmed the visual inspection of results. Compared to humans, the non-mentalistic models produced the largest errors (RMSE > 0.376 for all cases). Among the semi-mentalistic models, the *True belief* model provided an error comparable to the non-mentalistic models (RMSE = 0.367), the *Uncertainty rep.* model's error was about half that of the *True belief* model (RMSE = 0.174), and the partial representation had about half the error of the *Uncertainty rep.* (RMSE = 0.092) model. Finally, the human ToM model (*Ignorance rep.*) showed the lowest error (RMSE = 0.077).

Comparing the models to NHP data showed a different pattern of results. The largest discrepancy was against the human-like ToM model

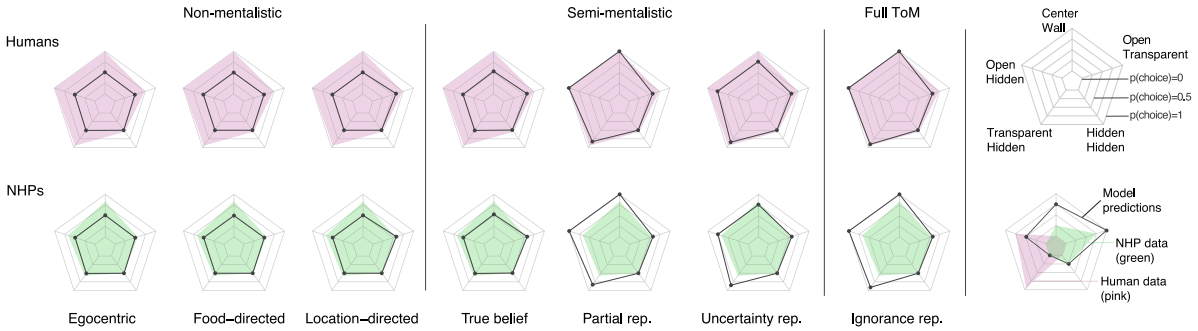

Fig. 3. Quantitative model predictions assuming full reliance. Each radar plot shows model-predicted effect sizes (black lines) compared to empirical data (human in pink, top row; NHP in green, bottom row), for each of the five paradigms (one per corner). Values represent the probability of choosing hidden food for the directional paradigms and the probability of choosing the right-most food in control paradigms. A tighter fit of the black line to the colored region indicates better model-data alignment. For the Hidden-Hidden paradigm, the NHP study reported a null result but not the exact effect size (Hare et al., 2000), so we set it to 50% for this visualization.

(*Ignorance representation*; RMSE = 0.225), followed by the *Partial rep.* model (RMSE = 0.203). The next three worst fits were the models in the non-mentalistic family (all RMSE > 0.167). Finally, the best two models were in the semi-mentalistic family: *True belief* (RMSE = 0.156) and *Uncertainty rep.* (RMSE = 0.134).

Together, these results suggest that the *Ignorance representation* model best fits human behavior, whereas the semi-mentalistic *True belief* and *Uncertainty rep.* models best fit NHP behavior. This is because the non-mentalistic models do not capture NHP behavior, and the human-like full ToM model predicts effect sizes that are much stronger than those documented.

2.2. According to each model, how often do NHPs rely on their social representations?

Our analyses so far have focused on understanding each model's core performance, which clarifies how we would expect NHPs (or humans) to behave if every individual always used these representations in every experiment. Having characterized how these models perform, we are ready for our key analysis of interest: formally estimating how much NHPs *rely* on their social representations, according to different theories. Our goal is to reveal the implicit assumption each account makes about how frequently its representations must be used in order to explain the data. This assumption is usually left unstated, but here, our computational framework allows us to extract it. Specifically, we accomplish this by searching over the space of possible reliance values and finding the one that maximizes each model's potential to explain NHP data. This approach uncovers each theory's implicit commitment to how frequently NHPs rely on their social representations to guide their behavior.

To find the value of reliance that best explains behavior, it is necessary to look beyond point estimates of the effect sizes and instead consider the full distribution of trial-level successes and failures. We approached this by finding the reliance parameter that maximizes each model's probability of generating the data, $P(D|M)$. Unfortunately, choice data were not always recoverable in the eleven NHP experiments we considered (Table 1), forcing us to limit our analysis to seven experiments and to drop the hidden-hidden paradigm. See SM Section 4.1 for details. We then calculated the probability of each model generating the behavioral choice data (i.e., the model's explanatory power), testing reliance values between 0 and 1 by intervals of 0.01 (see SM Figure S1). We conducted this analysis twice, once fitting reliance to explain human data, and again fitting reliance to explain NHP data.

We begin by evaluating the model fits against human judgments to confirm the validity of our analysis. Here, given the task's simplicity, we should expect that the best fit model is the human-like ToM one with a high reliance parameter. Our analysis confirmed this (Fig. 4A): The best model for explaining human data was the *Ignorance rep.* model positing reliance $r = 0.99$ ($CI_{95\%} = [0.97, 1.00]$; obtained through a nonparametric bootstrap, see SM Section 4.2). Because adult humans are independently known to readily deploy a full, human-like ToM in comparable tasks, recovering exactly this model and reliance confirms that our analysis works as intended.

Among the semi-mentalistic family, all three models maximize their fit to human data by positing high degrees of reliance ($r \geq 0.99$ in all cases; all $CI_{95\%}$ within $[0.99, 1.00]$), but they do not capture human data as well as the *Ignorance rep.* model. The *Partial rep.* was the closest (Partial: $LL = -36.3$, Ignorance: $LL = -25.2$). By contrast, the *Uncertainty rep.* and *True belief* models provided substantially lower fits (Uncertainty: $LL = -130.9$, True belief: $LL = -364.5$), showing that it is, in principle, possible for these paradigms to distinguish between subtly different accounts. Finally, the models in the non-mentalistic family all showed the lowest explanatory power. Indeed, the *Food-directed* model was such a poor fit to human judgments that its best explanatory power was reached under reliance $r = 0$, revealing that even the smallest use of this model negatively affected model fit.

To better interpret how well the models explain the data, we also calculated the highest explanatory power any model could achieve (the best-possible in-sample fit and log-likelihood; see SM Section 3.7 for a cross-validation analysis). This revealed a $LL = -10.0$ ceiling on explanatory power (depicted in Fig. 4A by the horizontal dotted line), measuring how difficult this dataset is to explain. Comparing the explanatory power of the *Ignorance rep.* model with this ceiling ($\Delta LL = -15.2$) reveals room for improvement (discussed in SM Section 2.5).

Having validated our approach against human data, we next applied the same approach to reveal the implicit reliance each theory must posit to explain NHP behavior (Fig. 4B). As foreshadowed by examining the models under full reliance, the non-mentalistic models showed the worst fit, and the *True belief* model had explanatory power nearly as low, confirming that our approach can still rule out poorly-fitting models even when they are free to posit any degree of reliance.

The three most complex mentalistic models (*Ignorance rep.*, *Partial rep.*, and *Uncertainty rep.*) all achieved similarly high explanatory power ($LL = -27.2$, -26.1, and -29.2, respectively). Because our approach reconstructs both the model used and its reliance from the same data, a rich representation used rarely and a simpler representation used often produce very similar fits. Consistent with this, the bootstrapped clouds of these three models occupy overlapping regions (Fig. 4B), and a model and parameter recovery simulation confirmed that they are confusable with one another in this reliance range at the current NHP sample sizes (SM Section 3.6). The data is therefore insufficient to distinguish between these representations.

Instead, this analysis uncovers how much reliance each account must posit, and here the accounts differ sharply. The human-like *Ignorance rep.* model predicts effect sizes much stronger than those found

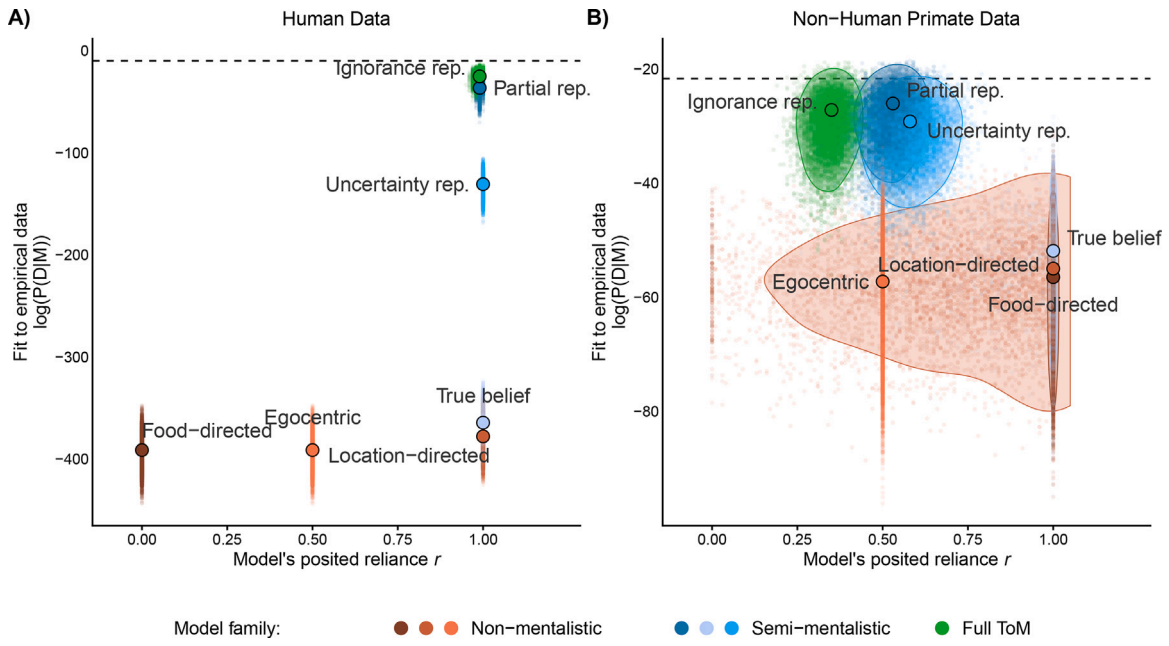

Fig. 4. Fitting each model's reliance to explain the empirical data for humans (Panel A) and non-human primates (Panel B), with bootstrapped uncertainty around each fit. Each plot shows the reliance parameter r (x-axis) that maximizes explanatory power (y-axis, in log-space); models higher on the y-axis have higher explanatory power; models higher assume more subjects use the posited representations. Large outlined points show each model's best-fitting (reliance, explanatory power) estimate. Small points show the fit recovered from a nonparametric bootstrap over the experimental trials (see SM Section 4.2); the shaded region for each model is the area containing 90% of these resampled fits. Color indicates each model's family (non-mentalistic, semi-mentalistic, or full Theory of Mind; see legend), and the reliance-invariant *Egocentric* model is plotted at a representative $r = 0.5$. In (B), the wide, diffuse cloud for *Location-directed* indicates that its reliance is not identifiable from the NHP data. Because it predicts close to chance on every paradigm, its explanatory power is essentially flat across reliance (see SM Section 1.3.1). The dashed horizontal line depicts the ceiling on each dataset's explanatory power. The two panels' y-axes differ in scale, since the human and NHP datasets contain different amounts of data.

in NHP behavior, so it can match the weaker observed effects only at a surprisingly low reliance: $r = 0.35$ ($CI_{95\%} = [0.26, 0.43]$). The simpler *Partial rep.* and *Uncertainty rep.* require more, positing $r = 0.53$ ($CI_{95\%} = [0.43, 0.62]$) and $r = 0.58$ ($CI_{95\%} = [0.45, 0.71]$), respectively (see SM Section 3 for a paradigm-by-paradigm walk-through). This reveals a previously hidden assumption in the debate over whether NHPs have a human-like ToM: if NHPs are using human-like representations in these tasks, they must only use them about 35% of the time, whereas if they are using a simpler type of ToM, they must use it about 55% of the time. This reliance gap is the main contribution of our analysis: it reveals that an account positing full human-like representations of ignorance must commit to a lower level of reliance than accounts positing simpler mentalistic representations in order to explain NHP behavior on visual perspective-taking tasks.

Finally, comparing the explanatory power of these three best models to the ceiling of how well this dataset could, in theory, be explained ($LL = -21.7$) shows that these models still leave some room for improvement (*Partial rep.*: $\Delta LL = -4.4$, $CI_{95\%} = [-13.0, -1.2]$; *Uncertainty rep.*: $\Delta LL = -7.5$, $CI_{95\%} = [-17.6, -2.7]$; and *Ignorance rep.*: $\Delta LL = -5.5$, $CI_{95\%} = [-14.6, -1.7]$). This leaves open the possibility that other mental representations may better explain NHP behavior on visual perspective-taking tasks.

2.3. A final check: could a different softmax temperature rescue simpler accounts?

Throughout, we fixed the softmax temperature parameter at $\tau = 0.05$. This is standard practice so that no single inconsistent choice can drive a model's likelihood to zero, but small enough that the policies remain near-optimal. However, because τ captures how sensitively an agent will respond to small differences in value, it determines how much small incidental features of a paradigm's geometry matter. This makes it possible that a model that fails at $\tau = 0.05$ might succeed at other values. We therefore conducted one final analysis where we refit each model to empirical data (Fig. 5A), and the recovered reliance (Fig. 5B) as a function of different τ values.

This analysis overall revealed that our exact conclusions are largely robust to τ parameter variations (Fig. 5). Can the non-mentalistic accounts or the *True belief* models be rescued at all? There is a parameter setting under which the *True belief* model recovers: $0.01 \leq \tau \leq 0.04$. Recall that τ, the softmax temperature, controls the sensitivity of choices to small differences in the value of options, which here amount to small differences in path length. When τ is in this range, choices become hypersensitive to these differences, allowing the *True belief* model to exploit a potential path-length confound in these paradigms. In the Center-Wall paradigm, the food behind the barrier appears to be marginally farther from the competitor, and in the Open-Hidden paradigm, the path to the food behind the barrier is longer because the competitor must navigate around it. At small τ the *True belief* model exploits such asymmetries in path lengths to strongly predict that the competitor will choose the food that is easier for it to reach, and so predicts that the subject will take the other, approximately reproducing NHPs' directional preferences on these paradigms. In this narrow window, the *True belief* model's fit rivals the best mentalistic models, peaking at $\tau = 0.03$ ($LL = -25.6$; best-fitting reliance $r = 1.00$, $CI_{95\%} = [0.88, 1.00]$; Fig. 5A).

Rather than rescuing the account, this exposes a hidden commitment: to maintain that NHP behavior reflects true-belief attribution, one must posit that the NHPs' preferential behavior in these particular studies results entirely from sensitivity to these incidental path-length differences. So construed, the *True belief* account becomes fundamentally about path-length sensitivities rather than representations of the competitor's mind. To our knowledge, no account of NHP social cognition has actually made the claim that hypersensitivity to path-length differences drives NHP behavior on these tasks. Furthermore, crediting *True belief*'s narrow peak requires converting τ from a noise parameter

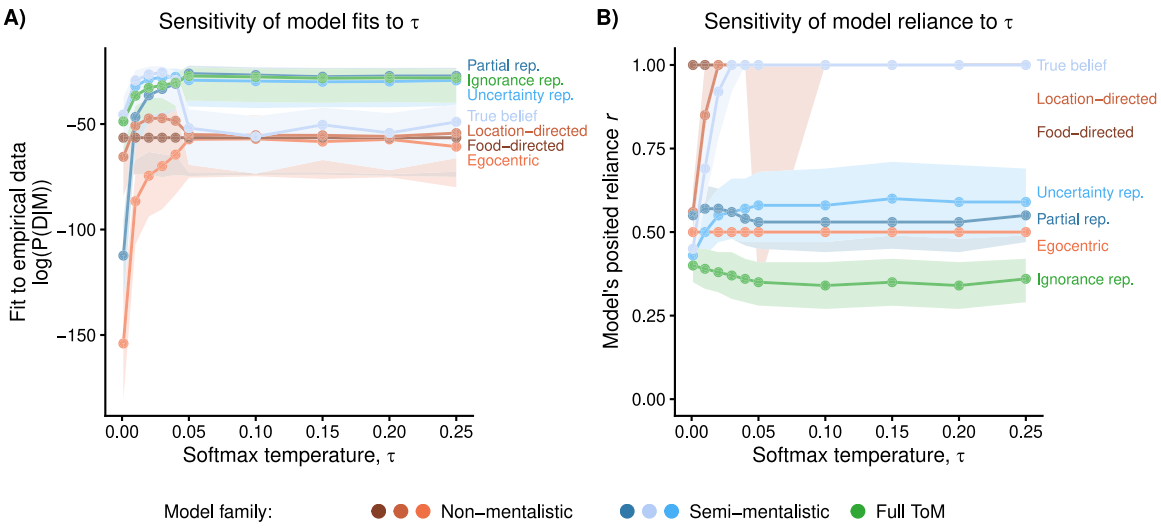

Fig. 5. Robustness of the analysis to the softmax temperature τ (x-axis). At each value of τ, each model's reliance r is refit to best explain the NHP data. **(A)** The explanatory power (fit) of each model. The *True belief* model shows an isolated peak at $\tau = 0.03$, whereas the leading mentalistic models are stable for $\tau \geq 0.05$. **(B)** The corresponding best reliance r for each model, which is stable across τ: the reliance gap between the accounts does not depend on the value of τ. The wide reliance ribbon on the *Location-directed* model at $\tau = 0.05$ recapitulates the wide cloud for that same model in Fig. 4B; see SM Section 1.3.1 for details.

common to all models into a substantive claim about how sensitively behavior tracks small differences in path length. Finally, the *True belief* account predicts identical behavior on the Open-Hidden test and the Open-Transparent control, and therefore cannot capture the behavioral difference NHPs exhibit between them.

Because the *Ignorance rep.*, *Partial rep.*, and *Uncertainty rep.* models' behavior is driven by their representations, not coincidences in path length, their fit does not hinge on τ. The reliance each account must posit, and the gap between them, remains unchanged across τ (Fig. 5B), except at the smallest values of τ, where the near-deterministic policy tolerates no noise in behavior and every model fits poorly. There is no escape from the central result: explaining NHP behavior requires accepting a tradeoff between the richness of the mentalistic representations and their reliance.

3. General discussion

3.1. Summary of findings

Here, we introduced a computational framework for comparative social cognition and applied it to visual perspective-taking as a case study. We implemented candidate theories of non-human primate (NHP) social representations as computational models and tested them on five paradigms from the perspective-taking literature. Using these models to derive the effect sizes that each theory predicts, we found that the models with the most complex representations best fit human behavior, while a variety of models including those with simpler representations of other minds best fit NHP behavior.

We then formalized a notion of *reliance* to model how frequently subjects would have to use their posited representations for each account to explain the observed data, and fit this degree of reliance for every model. The account that best fit human perspective-taking data is one where people use a complete Theory of Mind almost all the time; capturing NHP behavior, by contrast, required reducing reliance across all models. The most human-like model of Theory of Mind could explain primate behavior by assuming that NHPs only use these representations about one-third of the time, so positing that NHPs have human-like Theory of Mind comes with the implication that they only rarely use it; two other mentalistic accounts, positing simpler representations of other minds (*Partial representation* and *Uncertainty representation*; Table 2), capture the data at reliance just over 50%. Throughout, the specific models are a means rather than the central claim: what they provide is a way to make explicit the reliance each account must assume to explain NHP behavior.

3.2. Reliance distinguishes human and non-human primate theory of mind

Taken together, this approach converges with past investigations to validate the mentalistic nature of NHP visual perspective-taking, while also revealing a key difference between human and non-human primate social representations: the degree to which they are relied upon to guide behavior. Our work suggests that, if NHPs have a Theory of Mind of similar representational complexity to humans, then its usage is surprisingly limited compared to how readily human adults use their Theory of Mind, at least in visual perspective-taking tasks.

By contrast, human Theory of Mind is characterized not only by its cognitive richness but also by how often we use it in a wide range of contexts, even automatically in some cases (Kampis & Southgate, 2020). For instance, people will sometimes automatically track each other's knowledge from their visual perspective (Samson et al., 2010) and make automatic common ground inferences (Rubio-Fernández et al., 2019). Evidence of this kind of automaticity in non-human great apes is weaker, although some studies have found that the contents of an observer's beliefs (true or false) can influence their search behavior, as if automatic perspective-taking affects their search (Lurz et al., 2022; Padberg et al., 2025).

The difference in reliance raises the question of the comparability of the perspective-taking tasks that the NHPs and humans performed. The human and NHP tasks differed in several respects. Most of these differences, especially the higher stakes—competitive situations with known conspecifics (Hare & Tomasello, 2004), designed for ecological validity—and the live, physical format of the NHP task, should push NHPs toward higher motivation and reliance relative to online human participants (see SM Section 3.9 for further details.) The main exception is that NHP tasks may have imposed a time-constraint that the human tasks lacked. Under such conditions, humans may switch to non-mentalistic strategies to predict the behavior of other agents (Andrews et al., 2021), consistent with resource-rational accounts in which the deployment of costly mental-state reasoning is calibrated to its expected benefits. If so, the difference in reliance between humans and NHPs may be somewhat smaller than our work reports; nonetheless, its magnitude makes it clear that NHPs deploy their representations far less

often than humans. We hope that our work enables future discussions of comparative Theory of Mind to achieve more nuance by incorporating claims related to both representations and reliance.

In our current framework, reliance is a broad construct that captures many forces that may shape performance, such as motivation, the "effort" required to deploy representations, and cognitive load (e.g., time pressure, uncertainty). Although our framework subsumes all these factors under a single reliance parameter, future work could use this construct to tease apart the different causes behind degrees of reliance.

3.3. Relation to prior work and methodological considerations

Our approach adds to a growing literature on using computational modeling to understand NHP social cognition (e.g., social learning, communication, and coordination; Bohn et al., 2022; Devaine et al., 2017; Nong, 2023). However, past work has primarily focused on components of NHP behavior where computational models of human behavior could be directly applied to NHPs. Our approach implements models distinct from those capturing human cognition and tests for NHP-unique representations that are distinct from full Theory of Mind. Moreover, our work also makes explicit the reliance assumptions that competing theories of NHP mental representation implicitly carry, and evaluates each theory's explanatory power under those assumptions.

As in most computational work, formalizing theories into computational models required us to set multiple parameters. Our work is no exception: the values of the cost of movement, the magnitude of the food reward, the future discount, and the softmax parameter shape model fits. Here we took a conservative approach: minimizing their influence when possible (as with future discount and food reward), integrating over plausible values when unsure (as with costs), and conducting robustness analyses when their influence mattered (as with the softmax parameter). However, these parameters inevitably affect the exact predicted effect sizes and the estimated reliance. We see this as revealing that factors often overlooked in theories of NHP social reasoning—such as movement cost or reward magnitude—might be critical for theory evaluation, and that empirical work should attend to and study these components directly. We hope future research can estimate baseline energy expenditure, food motivation, and future discount parameters in tasks so as to guide parameter settings and even account for individual variation.

3.4. Limitations

Our work has three primary limitations. The first is that we only focused on seven computational models of visual perspective-taking. While these models varied substantially in their degree of cognitive complexity, they do not cover the full space of conceptual theories. We hope that in future work, researchers developing theories of NHP social behavior can work with modelers to implement their theories. Our framework provides a standardized way to estimate how well a theory captures NHP data and how often NHPs rely on the representation in order for the theory to hold. From this standpoint, our computational models can serve as a set of benchmarks that can quickly reveal whether new theories have strong explanatory power. Moreover, our computational approach can indicate when competing theories make identical predictions and guide future study design. Crucially, some theories (such as the minimal mindreading account; Butterfill & Apperly, 2013) are put forward in order to explain an existing set of findings without resorting to decoupled mental state representations (i.e., without a full Theory of Mind). In situations where competing theories explain existing data equally well, our modeling approach can be extended to generate trial types where competing theories ought to make distinct predictions, which can help generate decisive future studies.

Second, our framework assumes that NHPs rely on a single social representation and fall back to an egocentric baseline when that representation is not used, rather than using a combination of multiple representations (e.g., $r_1 \cdot p_{\text{model}_1} + r_2 \cdot p_{\text{model}_2} + \cdots$). We did not pursue this here because doing so would require theoretical accounts of which representations can coherently co-occur in an NHP's reasoning and which cannot, and which will be fall-backs for which others. One useful template for this future work comes from overlapping waves theory (Siegler, 1996) in human cognitive development, which models children as holding multiple problem-solving strategies simultaneously, with the mixture of strategies in use shifting gradually over development rather than switching discretely from one to the next; an analogous approach in comparative work would model NHPs as holding multiple social representations simultaneously, with the coherence of particular combinations, and the degree of reliance on each theory as open and promising directions for future theoretical, computational, and empirical work.

One final limitation of our work is that we exclusively focused on studies conducted with catarrhine primates: chimpanzees and Tonkean macaques. We chose this level/unit of analysis for two reasons: first, the catarrhine group includes humans, thus making other catarrhine species most relevant for human-nonhuman primate comparisons; second, catarrhines typically perform well on visual perspective-taking tasks and are often not considered to vary in this capacity, making them easier to model/evaluate as a group (Lewis & Krupenye, 2022). Our modeling approach can easily be applied to a wider range of species in the future as more data becomes available since perspective-taking capacities could vary dramatically across the primate taxon. Indeed, initial evidence suggests that capuchin monkeys (a species of platyrrhine, more distantly related to humans relative to catarrhines) and ring-tailed lemurs (a species of strepsirrhine, even more distantly related to humans relative to catarrhines and platyrrhines) are more limited in their understanding of the relationship between perception and knowledge relative to chimpanzees and macaques (Bray et al., 2014; Devaine et al., 2017; Hare et al., 2003). As such, the computational models that were discarded here for their poor qualitative fit to the catarrhine data (e.g., *Location-directed* and *True Belief*) might prove more useful for characterizing the social cognitive representations of platyrrhines and strepsirrhines. Similarly, as more data become available for additional catarrhine species, our approach could be used for species-level analyses targeted to detect nuanced distinctions within catarrhines, providing further insight into the evolutionary origins of primate Theory of Mind.

3.5. Future directions

While our focus was on visual perspective-taking tasks, our general framework can be adapted to Theory of Mind's many other components, including gaze-following (Bettle & Rosati, 2019; Drayton & Santos, 2017; MacLean & Hare, 2012; Rosati et al., 2016; Tomasello et al., 1998), awareness representations based on an agent's past perceptions (Drayton & Santos, 2018; Hare, 2001; Horschler et al., 2019, 2021; Santos et al., 2006), and belief representations (Horschler et al., 2020a, 2020b; Kaminski et al., 2008; Kano et al., 2020; Krupenye et al., 2016; Martin & Santos, 2016). 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.

Extending our framework to other Theory of Mind abilities will also invite a finer treatment of reliance, perhaps decomposing it into the forces that may shape performance, such as motivation, the "effort" required to deploy representations, and cognitive load (e.g., time pressure, uncertainty). Future versions could capture paradigm-specific reliance (some experimental paradigms might increase NHPs' reliance on their available social representations; e.g., Krupenye et al. (2016)) and possibly even model switching (NHPs might use different representations in different tasks). Even within visual perspective-taking tasks, variations in the paradigms appear to affect success: ecologically-relevant, competitive contexts where food rewards are far enough apart

to enhance attention to physical costs (Hare, 2001) seem to better reveal NHP social cognitive abilities than tasks where rewards are close together (Povinelli & Eddy, 1996). These factors can be quantitatively tested using our reliance construct. In particular, deliberately manipulating factors the present experiments hold fixed, such as the competitor's dominance rank, the available response time, or the appeal of the food reward, would test whether reliance tracks the cost-benefit trade-offs predicted by resource-rational accounts (e.g., Quillien and Taylor-Davies (2026)). Thus, our theory-testing framework can help future researchers test not only the representational underpinnings of a given capacity but also the factors that impact specific task performance.

3.6. Reliance and the competence–performance distinction

Our construct of reliance is closely aligned with and formally instantiates the classic competence–performance distinction (Chomsky, 1965). This distinction has been central to debates about whether infants and young children lack belief/desire understanding (a competence limitation) or instead possess the relevant representations but fail to deploy them under the linguistic, inhibitory, or motivational task demands (a performance limitation; Apperly, 2012; Baillargeon et al., 2010; Bloom & German, 2000; Surian & Leslie, 1999). Our framework formalizes the same conceptual distinction, separating the cognitive representations from the degree to which they govern behavior (reliance). By quantifying reliance, our approach provides an avenue for evaluating not only how well candidate representations explain behavior, but when and why they drive behavior. Our work provides a shared quantitative vocabulary for comparing social representations and their deployment in NHPs across species and paradigms.

4. Conclusion

Understanding what non-human primates (NHPs) know about each other's minds has been a central goal in comparative cognitive science for over forty years. Progress has been challenging, however, because NHPs can fail to reveal their capacities in some tasks or succeed through simpler strategies in others.

Our work advances this debate by formalizing the different theories and estimating the degree of reliance each theory must posit to account for behavior. Our results show that both human-like and non-human-like representations can explain NHP behaviors, but each assumes different degrees of reliance on these representations. Consequently, regardless of which theory one favors, NHPs do not resemble humans in how frequently and pervasively we use our Theory of Mind. Perhaps, in NHP social environments, applying Theory of Mind is not often necessary, possibly because NHPs have alternative role-based systems for understanding behavior (Jara-Ettinger & Dunham, 2025). But this does not mean they lack mentalistic representations; on the contrary, our framework shows strong support for mentalistic representations. Thus, our findings are not deflationary but instead invite researchers to explore when NHPs use their Theory of Mind and when they rely on alternative non-mentalistic computations to solve complex social problems.

CRediT authorship contribution statement

Marlene D. Berke: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Daniel J. Horschler:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Amanda Royka:** Writing – review & editing. **Laurie R. Santos:** Conceptualization. **Julian Jara-Ettinger:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During revision of this work, the authors used Claude to improve the clarity, language, and readability of the manuscript and to assist in writing and checking analysis and figure-generation code. As part of these revisions, the authors also used Claude to draft portions of the Supplementary Materials, specifically SM Sections 1.3.1, 3.6, 3.7, 3.8, 3.9, and 4.2. After using this tool, the authors edited and reviewed the content and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Mel Andrews, Laura Schulz, and Max Siegel for helpful comments. This work was supported by NSF CAREER award BCS-2045778 awarded to JJE, NSF SBE Postdoctoral Research Fellowship 2104589 awarded to DJH, and the Lisa K. Yang Integrative Computational Neuroscience Fellowship awarded to MDB.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2026.106698.

Data availability

All data and code are available in our OSF repository (Berke et al., 2026): https://doi.org/10.17605/OSF.IO/QJN9M.

References

Andrews, K. (2012). *Do apes read minds?: Toward a new folk psychology*. MIT Press.
Andrews, K. (2018). Apes track false beliefs but might not understand them. *Learning & Behavior*, *46*, 3–4.
Andrews, K., Spaulding, S., & Westra, E. (2021). Introduction to folk psychology: Pluralistic approaches. *Synthese*, *199*(1), 1685–1700.
Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. http://dx.doi.org/10.1080/17470218.2012.676055.
Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. http://dx.doi.org/10.1016/j.tics.2009.12.006.
Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.
Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.
Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, *6*(5), 679–684.
Berke, M. D., Horschler, D. J., Royka, A., Santos, L. R., & Jara-Ettinger, J. (2026). What primates know about other minds and how often they use it: Models, analysis code, and data. URL https://doi.org/10.17605/OSF.IO/QJN9M. [dataset]. OSF.
Bettle, R., & Rosati, A. G. (2019). Flexible gaze-following in rhesus monkeys. *Animal Cognition*, *22*(5), 673–686.
Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31.
Bohn, M., Liebal, K., Oña, L., & Tessler, M. H. (2022). Great ape communication as contextual social inference: a computational modelling perspective. *Philosophical Transactions of the Royal Society B*, *377*(1859), Article 20210096.
Bräuer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal Cognition*, *10*(4), 439–448.
Bray, J., Krupenye, C., & Hare, B. (2014). Ring-tailed lemurs (lemur catta) exploit information about what others can see but not what they can hear. *Animal Cognition*, *17*(3), 735–744.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606–637.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192. http://dx.doi.org/10.1016/j.tics.2008.02.010.

Canteloup, C., Piraux, E., Poulin, N., & Meunier, H. (2016). Do Tonkean macaques (macaca tonkeana) perceive what conspecifics do and do not see? *PeerJ*, *4*, e1693.

Cheney, D. L., & Seyfarth, R. M. (2008). Baboon metaphysics: the evolution of a social mind. University of Chicago Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133.

Devaine, M., San-Galli, A., Trapanese, C., Bardino, G., Hano, C., Saint Jalme, M., Bouret, S., Masi, S., & Daunizeau, J. (2017). Reading wild minds: a computational assay of theory of mind sophistication across seven primate species. *PLoS Computational Biology*, *13*(11), Article e1005833.

Drayton, L. A., & Santos, L. R. (2017). Do rhesus macaques, macaca mulatta, understand what others know when gaze following?. *Animal Behaviour*, *134*, 193–199.

Drayton, L. A., & Santos, L. R. (2018). What do monkeys know about others' knowledge? *Cognition*, *170*, 201–208.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gordon, R. M. (1992). The simulation theory: Objections and misconceptions. *Mind & Language*, *7*, 11–34.

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

Halina, M. (2017). What apes know about seeing. In *The routledge handbook of philosophy of animal minds* (pp. 238–246). Routledge.

Hall, K., & Brosnan, S. F. (2017). Cooperation and deception in primates. *Infant Behavior and Development*, *48*, 38–44.

Hare, B. (2001). Can competitive paradigms increase the validity of experiments on primate social cognition? *Animal Cognition*, *4*(3–4), 269–280.

Hare, B., Addessi, E., Call, J., Tomasello, M., & Visalberghi, E. (2003). Do capuchin monkeys, cebus apella, know what conspecifics do and do not see? *Animal Behaviour*, *65*(1), 131–142.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*, 771–785.

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, *101*(3), 495–514.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, *68*(3), 571–581.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267–272.

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, *21*(1), 101–114.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*, 959–971.

Hoppitt, W. J., Brown, G. R., Kendal, R., Rendell, L., Thornton, A., Webster, M. M., & Laland, K. N. (2008). Lessons from animal teaching. *Trends in Ecology & Evolution*, *23*(9), 486–493.

Horschler, D. J., MacLean, E. L., & Santos, L. R. (2020a). Advancing gaze-based research on primate theory of mind. *Trends in Cognitive Sciences*, *24*(10), 778–779.

Horschler, D. J., MacLean, E. L., & Santos, L. R. (2020b). Do non-human primates really represent others' beliefs? *Trends in Cognitive Sciences*, *24*(8), 594–605.

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*, *190*, 72–80.

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2021). How do non-human primates represent others' awareness of where objects are hidden? *Cognition*, *212*, Article 104658.

Hutto, D. D. (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. MIT Press.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jara-Ettinger, J., Baker, C., Ullman, T., & Tenenbaum, J. B. (2024). Theory of mind and inverse planning. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Bayesian models of cognition: reverse engineering the mind* (pp. 341–368). Cambridge, MA: MIT Press.

Jara-Ettinger, J., & Dunham, Y. (2025). The institutional stance. *Behavioral and Brain Sciences*, http://dx.doi.org/10.1017/S0140525X25103427, Advance online publication.

Jara-Ettinger, J., Schulz, L., & Tenenbaum, J. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, Article 101334.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*(2), 224–234.

Kampis, D., & Southgate, V. (2020). Altercentric cognition: How others influence our cognitive processing. *Trends in Cognitive Sciences*, *24*(11), 945–959.

Kano, F., Call, J., & Krupenye, C. (2020). Primates pass dynamically social anticipatory-looking false-belief tests. *Trends in Cognitive Sciences*, *24*(10), 777–778.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), Article e1503.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114.

Lewis, L. S., & Krupenye, C. (2022). Theory of mind in nonhuman primates. In *Primate Cognitive Studies* (pp. 439–482). Cambridge University Press.

Lewis, L. S., Wessling, E. G., Kano, F., Stevens, J. M., Call, J., & Krupenye, C. (2023). Bonobos and chimpanzees remember familiar conspecifics for decades. *Proceedings of the National Academy of Sciences*, *120*(52), Article e2304903120.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS One*, *9*(3), Article e92160.

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*, *12*(1), 142–149.

Lurz, R. W., Kanet, S., & Krachun, C. (2014). Animal mindreading: A defense of optimistic agnosticism. *Mind & Language*, *29*(4), 428–454.

Lurz, R. W., Krachun, C., Mareno, M. C., & Hopkins, W. D. (2022). Do chimpanzees predict others' behavior by simulating their beliefs? *Animal Behavior and Cognition*, *9*(2), 153–175.

MacLean, E. L., & Hare, B. (2012). Bonobos and chimpanzees infer the target of another's attention. *Animal Behaviour*, *83*(2), 345–353.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, *20*(5), 375–382.

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology*, *120*(2), 154–162.

Muller, M. N., & Mitani, J. C. (2005). Conflict and cooperation in wild chimpanzees. *Advances in the Study of Behavior*, *35*, 275–331.

Musgrave, S., Morgan, D., Lonsdorf, E., Mundry, R., & Sanz, C. (2016). Tool transfers are a form of teaching among chimpanzees. *Scientific Reports*, *6*(1), 34783.

Noë, R., de Waal, F. B., & van Hooff, J. A. (1980). Types of dominance in a chimpanzee colony. *Folia Primatologica*, *34*(1–2), 90–110.

Nong, T. (2023). *Computational modeling applied to strategic social decision making and theory of mind across different species of primates* (Ph.D. thesis), Université Claude Bernard-Lyon I.

Padberg, M., Hanus, D., & Haun, D. (2025). Great apes show altercentric influences when confronted with conflicting beliefs. *Animal Behaviour*, *227*, Article 123304.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *362*(1480), 731–744.

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

Povinelli, D. J., & Eddy, T. J. (1996). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development*, *61*(3).

Quillien, T., & Taylor-Davies, M. (2026). Factive mindreading reflects the optimal use of limited cognitive resources. *Proceedings of the Royal Society B*, *293*(2070), Article 20251852.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Rochat, M. J., Serra, E., Fadiga, L., & Gallese, V. (2008). The evolution of social cognition: Goal familiarity shapes monkeys' action understanding. *Current Biology*, *18*(3), 227–232.

Rosati, A. G., Arre, A. M., Platt, M. L., & Santos, L. R. (2016). Rhesus monkeys show human-like changes in gaze following across the lifespan. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1830), Article 20160376.

Rosati, A. G., Santos, L. R., & Hare, B. (2010). Primate social cognition: Thirty years after Premack and Woodruff. In *Primate neuroethology* (pp. 117–143). Oxford University Press.

Royka, A., & Jara-Ettinger, J. (2021). Ignorance matters. *Behavioral and Brain Sciences*, *44*.

Royka, A., & Santos, L. R. (2025). Informing does not require attributing ignorance. *Proceedings of the National Academy of Sciences*, *122*(20), Article e2504155122.

Rubio-Fernández, P., Berke, M. D., & Jara-Ettinger, J. (2025). Tracking minds in communication. *Trends in Cognitive Sciences*, *29*(3), 269–281.

Rubio-Fernández, P., Mollica, F., Ali, M. O., & Gibson, E. (2019). How do you know that? Automatic belief inferences in passing conversation. *Cognition*, *193*, Article 104011.

Saldana, C., Claidière, N., Fagot, J., & Smith, K. (2022). Probability matching is not the default decision making strategy in human and non-human primates. *Scientific Reports*, *12*(1), 13092.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255.

Santos, L. R., Nissen, A. G., & Ferrugia, J. A. (2006). Rhesus monkeys, macaca mulatta, know what others can and cannot hear. *Animal Behaviour*, *71*(5), 1175–1181.

Siegler, R. S. (1996). *Emerging minds: the process of change in children's thinking*. New York: Oxford University Press.

Sigaud, O., & Buffet, O. (2013). *Markov decision processes in artificial intelligence*. John Wiley & Sons.

Surian, L., & Leslie, A. M. (1999). Competence and performance in false belief understanding: A comparison of autistic and normal 3-year-old children. *British Journal of Developmental Psychology*, *17*(1), 141–155. http://dx.doi.org/10.1348/026151099165203.

Tomasello, M., Call, J., & Hare, B. (1998). Five primate species follow the visual gaze of conspecifics. *Animal Behaviour*, *55*(4), 1063–1069.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691.

Townrow, L. A., & Krupenye, C. (2025). Bonobos point more for ignorant than knowledgeable social partners. *Proceedings of the National Academy of Sciences*, *122*(6), Article e2412450122.

Wang, Z., Davis, I., & Jara-Ettinger, J. (2025). Modeling other minds: A computational account of social cognition and its development. *Annual Review of Developmental Psychology*, *7*, 65–89.

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology*, *5*(7), Article e184.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245–275.

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

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240.