Factive mindreading reflects the optimal use of limited cognitive resources

Tadeg Quillien[1] and Max Taylor-Davies[2]

[1]Department of Psychology
[2]School of Informatics
[1,2]University of Edinburgh

March 31, 2025

Abstract

Predicting what other individuals will do is an important adaptive challenge for many organisms. Social prediction can be achieved by constructing a detailed model of the mental states of other agents, but this is computationally expensive. We argue that mindreaders can often bypass the need for constructing such a detailed model: they can keep track of the facts in their own world model that another agent also knows, instead of explicitly representing the content of the agent's world model. Using a simple computational approach, we find that this 'factive' mindreading strategy emerges as the optimal social prediction strategy for organisms with limited cognitive resources across a range of social ecologies. Factive mindreaders in our model behave like young human children and non-human primates: they successfully predict the behavior of knowledgable and ignorant agents, but fail to predict the behavior of agents with false and even accidentally true beliefs. Our results elucidate the computational principles underlying efficient social prediction, and provide a first-principles account for a range of empirical findings about human and non-human mindreading.

Keywords: theory of mind, knowledge, false belief, information theory, resource rationality, social cognition

1 Introduction

Predicting the behavior of other individuals is a key adaptive challenge for most organisms. The challenge of social prediction has been a key driver of the evolution of Theory of Mind, or 'mindreading': the ability to represent the latent mental states of others. This capacity has been extensively studied across multiple domains, including its evolutionary origins, development in children, neural mechanisms, and conceptual structure [1–7].

Researchers have started to develop formal theories of mindreading at the computational level [e.g. 6, 8, 9]. These theories typically use normative models: they assume that mindreaders are making sophisticated, Bayesian computations over a causal model of the mental states of other agents. This approach has been successful for explaining the successes of mindreading in human adults and children across many tasks [6, 9, 10]. However, a normative approach is less well suited to explaining the patterns of systematic mistakes that participants—especially younger children and non-human animals—make in mindreading tasks [1, 11]. In this paper, we develop a computational approach that can shed light on both the successes and limitations of mindreading. Like existing models, we take a normative approach, asking how a well-designed cognitive system for mindreading would work. However, we also consider how this cognitive system would deal with limitations in computational resources. From this perspective, systematic mistakes can be understood as resulting from cognitive 'shortcuts' that save computation [12].

To understand existing approaches to mindreading, it is useful to understand the notion of an internal mental *model*. Cognitive scientists have long argued that agents can better interact with their environment if they construct an approximate model of that environment [13–15]. Applied to mindreading, this thesis has two related implications. First, agents often behave as if they had a model of their environment. Second, and as a consequence, a good way to predict the behavior of another agent is to construct a model of that agent's own internal world model. Accordingly, theories of mindreading emphasize the importance of meta-representation, the ability to represent others' representation of the world [6, 16, 17]. Observers that construct meta-representations are modeling the way another agent models its environment. For example, if Alice and Bob are in a room, Bob's world model might contain the information that 'there is an apple on the table', and Alice can meta-represent that 'Bob thinks: "there is an apple on the table"', see Figure 1A.

This normative ideal creates a puzzling challenge: meta-representation can be extremely costly in computational terms, especially in complex or densely populated environments. Consider just the memory demands: storing another individual's complete model of the world could in principle require as much memory as your own model of the world. Maintaining even partial models of the minds of everyone one knows would consume a substantial portion of the cognitive resources required for other critical tasks.

How do organisms achieve efficient social prediction despite these apparently prohibitive computational costs? Our key insight is that social prediction can often bypass the need for meta-representation. When individuals share the same physical environment, this creates substantial overlap in their mental models. Efficient mindreaders can leverage this redundancy to represent others' mental states more economically.

Consider again Alice and Bob, in the same room: they share much of their knowledge about their environment, such as seeing an apple on the table. Alice could in principle represent the fact that 'there is an apple on the table' twice: once in her own model of the world, and once in her model of Bob's world model (her meta-representation), see Figure 1 Left. But Alice can instead use a simpler strategy: she can store 'there is an apple on the table' in her own world model, and add a simple tag noting that Bob also has access to this fact (Figure 1 Right; [18]). This simpler strategy is called 'factive' because it represents relations between the other agent and true facts about the world [19, 20]. Factive mindreading is less flexible than meta-representation, because it does not allow Alice to model Bob as having a different belief than hers. But by avoiding duplicate representations of the shared environment, it comes at a significantly lower computational cost.

Here we show that factive mindreading can be an optimal strategy for social prediction in organisms with limited computational resources. Our work provides an evolutionary explanation for an emerging body of empirical work that suggests that both human and non-human animals often engage in factive mindreading [1, 18, 19, 21, 22]. For example, both human and non-human primates struggle to predict the behavior of individuals with false and accidentally true beliefs, but are able to predict the behavior of other individuals on the basis of what they know [23].

Formally, we conduct a *resource-rational analysis* of mindreading in a simple model of social prediction. In a resource-rational analysis, researchers seek to derive the optimal policy for solving an information-processing problem, under the constraint that this policy has to be executed by an agent with limited computational resources [12, 24, 25]. Here we consider a large space of possible policies for social prediction, and find the policies that optimize predictive performance under computational resource constraints. This process allows us to study social prediction policies that have been 'designed' by a normative optimization process, rather than hand-coded by a researcher. We find that resource-rational policies instantiate factive mindreading across a wide range of social ecologies.

We operationalize computational limitations in information-theoretic terms, as a bound on how much information the observer is able to extract from the environment. The advantage of this approach is that it allows us to remain agnostic with respect to particular implementation or substrate details—since information-theoretic constraints can be interpreted in multiple ways, such as limitations on inference or memory [26].

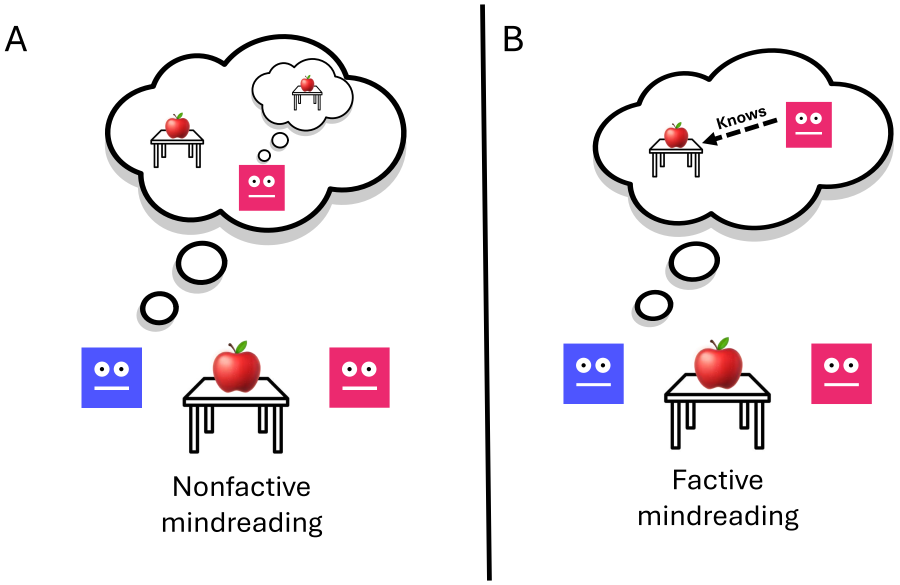

Figure 1: Difference between nonfactive and factive mindreading (adapted from [18]). **A**: The mindreader (blue) represents a fact (the apple is on the table) in its primary representation of the world, and also represents the other agent (pink) as representing that there is an apple on the table (a meta-representation). **B**: A factive mindreader simply tracks whether the other individual has epistemic access to a fact in the mindreader's world model.

Information-theoretic principles have been widely used in models of resource-rational cognition [27–42]. They offer a principled way to model cognitive resource limitations in the abstract, without making strong assumptions about cognitive architecture [26].

Our computational approach allows us to model social cognition without relying on folk-psychological concepts. We will still use some folk-psychological language *for ease of exposition*. Specifically, for convenience we say that factive mindreaders represent what other agents *know* instead of what they *believe* [23]. For our purposes, 'knowledge' denotes two important properties: agents can only know things that are true, and accidentally true beliefs do not count as knowledge [20, 43]. In contrast, the content of an agent's belief is a proposition like 'the apple is on the table'. Representing belief is more computationally costly than representing knowledge, but it comes with the flexibility of allowing the mindreader to faithfully represent agents whose world model differs from their own (e.g. in the case of false beliefs).

2 Modeling framework

We consider an *observer* who has to predict the behavior Y of an *agent*—for example the observer must predict where the agent will look for an apple. The observer has access to a stream of data $\vec{X}$ from the world, some of which is relevant to predicting the agent's behavior (information about the agent's location, gaze direction, etc). An observer with limited cognitive resources cannot process in detail all the information contained in the incoming sensory data, so they need to construct a compressed representation Z, that they will then use to predict Y. Ideally, Z extracts the information in $\vec{X}$ that is most relevant to the task of predicting the other individual's behavior (Figure 2 lower-right).

This problem can be formalized using the information bottleneck [44], a framework closely related to rate-distortion theory [45]. In an information bottleneck problem, we seek to construct an optimal encoder from $\vec{X}$ to Z. Formally, an encoder is a conditional probability distribution $q(z|\vec{x})$ that specifies the probability that the observer will form the representation $Z = z$ given that the state of the world is $\vec{X} = \vec{x}$, for all possible values of $\vec{x}$ and z.

The computational capacity of the observer is defined as an upper bound on the mutual information between $\vec{X}$ and Z:

$$I(\vec{X}; Z) = \sum_{\vec{x},z} \mathrm{Pr}(\vec{x}, z) \log \frac{\mathrm{Pr}(\vec{x}, z)}{\mathrm{Pr}(\vec{x})\mathrm{Pr}(z)} \tag{1}$$

where $Pr(\vec{x}, z) = q(z|\vec{x})Pr(\vec{x})$. Intuitively, this value quantifies the amount of information that compressed representation Z can 'preserve' about the input data $\vec{X}$. Given this upper bound on mutual information, the goal is to find an encoder that, on average, yields the representation Z that is most useful for predicting the agent behavior Y.

Crucially, we assume that the observer also has a representation S of the state of the physical world, because this representation is generally useful even outside of the context of social prediction. For example, the observer tracks the true location of the apple because they may want to eat it themselves. We assume that the information-theoretic costs of building representation S have already been paid by the observer, so that it can effectively be re-used for free in the social prediction task. We can then re-frame the task as that of predicting Y from Z and S *jointly* (see Figure 2 lower-right), with the usefulness of Z quantified as the additional predictive power that it gives the observer about Y, given that the observer already represents S. This quantity is operationalized as a conditional mutual information:

$$I(Y; Z|S) = I(Y; Z, S) - I(Y; S) \tag{2}$$

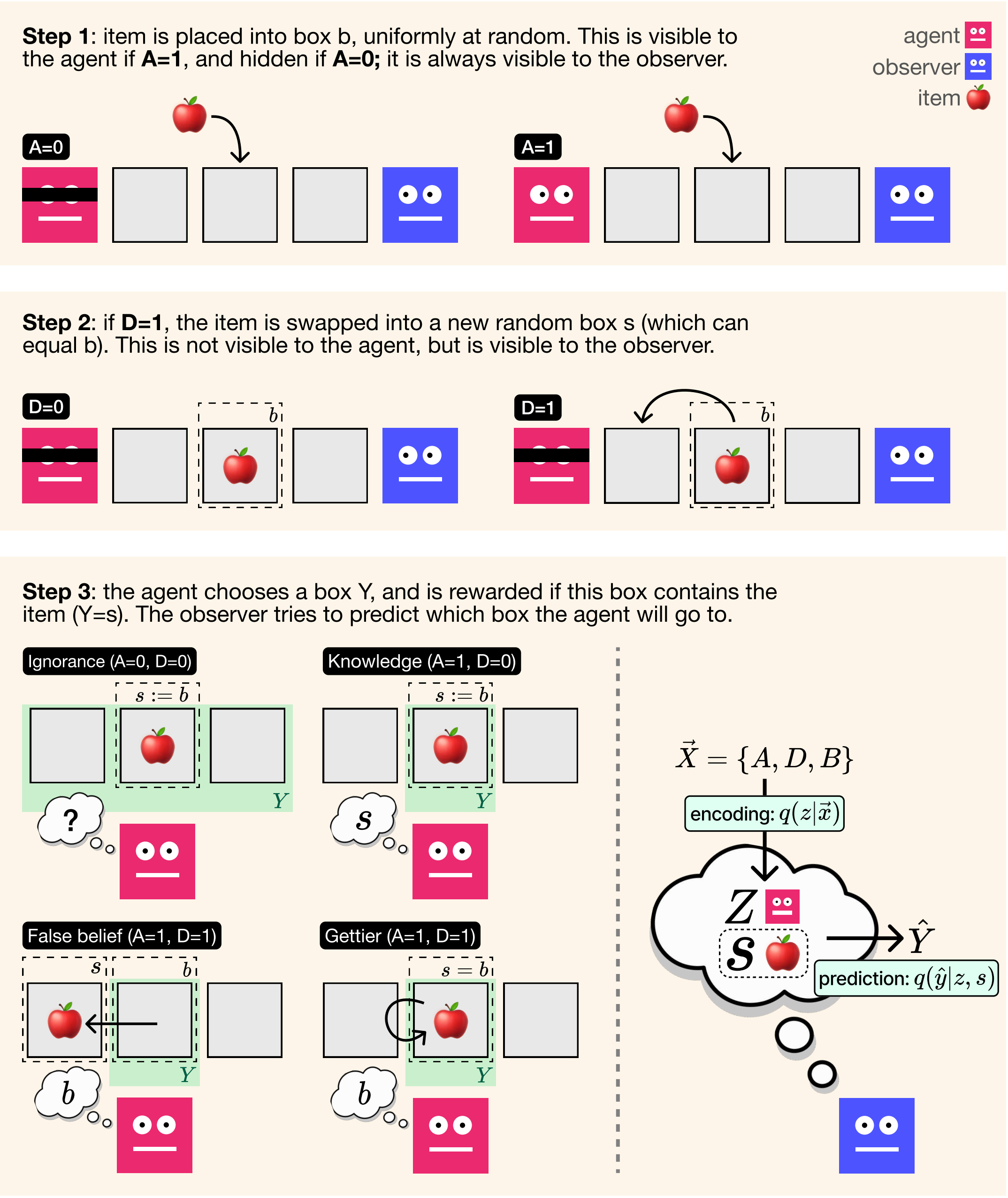
Step 1: item is placed into box b, uniformly at random. This is visible to the agent if A=1, and hidden if A=0; it is always visible to the observer.
agent
observer
item
A=0
A=1
Step 2: if D=1, the item is swapped into a new random box s (which can equal b). This is not visible to the agent, but is visible to the observer.
D=0
D=1
b
b
Step 3: the agent chooses a box Y, and is rewarded if this box contains the item (Y=s). The observer tries to predict which box the agent will go to.
Ignorance (A=0, D=0)
s := b
Y
?
Knowledge (A=1, D=0)
s := b
Y
s
False belief (A=1, D=1)
s
b
Y
b
Gettier (A=1, D=1)
s = b
Y
b
$\vec{X} = \{A, D, B\}$
encoding: $q(z|\vec{x})$
Z
S
$\hat{Y}$
prediction: $q(\hat{y}|z, s)$

Figure 2: Social prediction task and theoretical framework. In Step 3, green shading indicates the box(es) that the agent is most likely to go to in each case; thought bubbles represent where the agent thinks the item is. Lower right: information bottleneck model. $\vec{X}$ represents information in the world relevant to mindreading, such as what the agent (pink) can and cannot see. The observer (blue) constructs a compressed representation Z on the basis of $\vec{X}$, and also has access to additional representation S which reflects the world state (i.e. the true item location). The observer then uses Z and S to make a prediction $\hat{Y}$ about the agent's box choice Y.

In sum, we are looking for the optimal encoder:

$$q_C(z|\vec{x})^{\star} = \underset{q}{\text{arg max}}\ I(Y;Z|S) \qquad (3)$$
$$\text{subject to } I(\vec{X};Z) \leq C$$

where C is the upper bound on the amount of information the observer can extract from $\vec{X}$. Given the conditionalization on S, our problem is an instance of the *conditional* information bottleneck, and we solve it using the algorithm derived in [46], see Methods.

Importantly, we are not arguing that resource-rational mindreaders are solving Equation 3 themselves. Instead, resource-rational analysis takes the perspective that the constrained optimization problem has been approximately solved over time by evolutionary, developmental or learning processes, and that the observer is simply executing the resulting policy [12]. For simplicity we focus on the cognitive costs involved in constructing representation Z, but not in the costs involved in deriving a prediction from Z (following e.g. [33, 47]). Therefore we assume that the agent predicts behavior Y with the Bayes-optimal decoder $q(\hat{y}|z)$.

2.1 Task

We study the resource-rational mindreading problem in a simple task in which the observer must predict the behavior of an agent.

2.1.1 Agent's task.

The agent faces N opaque boxes. One of these boxes B is selected uniformly at random, and a valuable item (such as the apple in our earlier example) is placed into box B. The agent will have to choose a box and gets reward r if it picks the box containing the item, and 0 otherwise, see Figure 2.

With some probability $Pr(A)$, the agent has perceptual access and can see in which box the item is initially being placed (i.e. it can see which box is selected as B). Otherwise (with probability $1 - Pr(A)$), the agent is ignorant and receives no information about the item's location.

With probability $Pr(D)$, we then switch the item to a box S, selected uniformly at random (this can be the original box B), *always* outside of the agent's awareness. This 'Deceiver' event ($D = 1$) implies that any belief that the agent has formed might now be false. We use S to denote the final location of the item; if the item did not get switched we simply have $S = B$.

Following recent computational approaches to mindreading [6, 8, 9, 48], we assume that the agent is approximately rational and seeks to maximize expected reward given the information it has access to (see Supplementary Information). This means that the agent chooses a box uniformly at random if it did not see where the item was placed ($A = 0$). Otherwise ($A = 1$), it goes to the box where it last saw the item (box b), although it may sometimes choose a different box by mistake.

2.1.2 Observer's task.

The observer already has a representation of current location S of the item, and can extract the value of A, D and B as input data from the environment; i.e. we have $\vec{X} = \{A, D, B\}$. The observer's goal is to accurately predict where the agent will go, that is, to accurately estimate the probability of each choice.

This simple setting allows us to explore different situations traditionally studied in Theory of Mind research, including tasks where the observer must predict the behavior of an agent with knowledge ($A = 1$, $D = 0$), false belief ($A = 1$, $D = 1$, $s \neq b$), accidentally true belief ($A = 1$, $D = 1$, $s = b$), and ignorance ($A = 0$). Intuitively, variable B represents the 'content' of the agent's belief (assuming that $A = 1$), while A and D determine whether the agent knows the item location (specifically, the agent has knowledge if $A = 1$ and $D = 0$).

Below we derive the resource-rational observer policies for this task, using the framework outlined in the previous section, and investigate their properties. We call the combination of parameters $Pr(A)$, $Pr(D)$ and N the *social ecology*; in addition to these, we also vary the computational capacity C of the observer. In a given simulation, the values of parameters $Pr(A)$, $Pr(D)$, N and C are fixed, and the resource-rational policy is optimized for its expected performance across all possible settings of A, D, B and S (the probability of each setting is determined by the social ecology). Note that different social ecologies can in principle favor different resource-rational policies; in this sense resource-rational policies are *ecologically rational* [49, 50].

We predict that an observer that can only dedicate limited resources to constructing Z should focus these limited resources on encoding information that is least likely to be redundant with S. From this perspective, encoding the content B of the agent's belief ('the apple is in box 3') can be wasteful, because this information is typically already in the observer's own representation S. Instead, the observer can encode the value of A and D: whether the agent 'knows' the location of the apple. We therefore define as *factive* a policy that i) extracts little or no information about B, ii) extracts relatively more information about A and D. We measure the information extracted about a variable as the mutual information between the variable and compressed representation Z. Code for implementing our model is available on the Open Science Framework.

3 Results

We find that factive mindreading emerges as the optimal cognitive strategy across a large portion of the parameter space, see Figure 3B. In many social ecologies, resource-rational observers with low computational capacity extract substantially more information about the knowledge-relevant variables A and D than the belief-relevant variable B. Figure 3A illustrates this pattern for one example social ecology: observers with low computational capacity extract information about A and D but not about B, which is only represented by observers above a certain capacity threshold.

We can explore the representations formed by factive observers by visualizing the mapping from $\vec{X}$ to Z to $\hat{Y}$ in observers that extract no information about B. A detailed example is given in the Supplementary Information (Figure S3). Here we give a high-level overview of this content, showing an idealized depiction of the mapping performed by factive vs meta-representational observers (Figure 4). We find that factive observers have a representation Z that can be in only two possible states: the observer either represents the agent as being Ignorant (whenever $A = 0$ or $D = 1$) or Knowledgeable (whenever $A = 1$ and $D = 0$). Correspondingly, the observer predicts that an Ignorant agent might go toward any box, and predicts that a Knowledgeable agent will go to box S (the box that actually contains the item). We can see that $I(B; Z) = 0$ because the representation Z formed by the observer does not depend on B: for example an agent with $A = 1$, $D = 0$ is assigned to the same representation (Knowledge) regardless of the value of B.

In contrast, in social ecologies with high values of $Pr(D)$ and $Pr(A)$, or for observers with high computational capacity, the resource-rational policy is closer to a meta-representational policy. The lower panel on Figure 4 is an idealized depiction of a meta-representational policy. The observer represents the content of the agent's beliefs, like the belief that the item is in box 1. The representation Z extracts all the available information about B ($I(B; Z)$ is high), and the observer does not use its own representation of the state of the world S.

Figure 3B shows the prevalence of factive mindreading across social ecologies. Factive mindreading is especially prevalent for low values of $Pr(D)$ and low-to-intermediate values of $Pr(A)$. The likelihood of a false belief is equal to $Pr(A)Pr(D)$, and factive policies make sub-optimal predictions when the agent has a false belief, since in this case the item location S is not sufficient to predict what the agent will do. Factive mindreading is also more prevalent with increasing N (see SI), because the information-

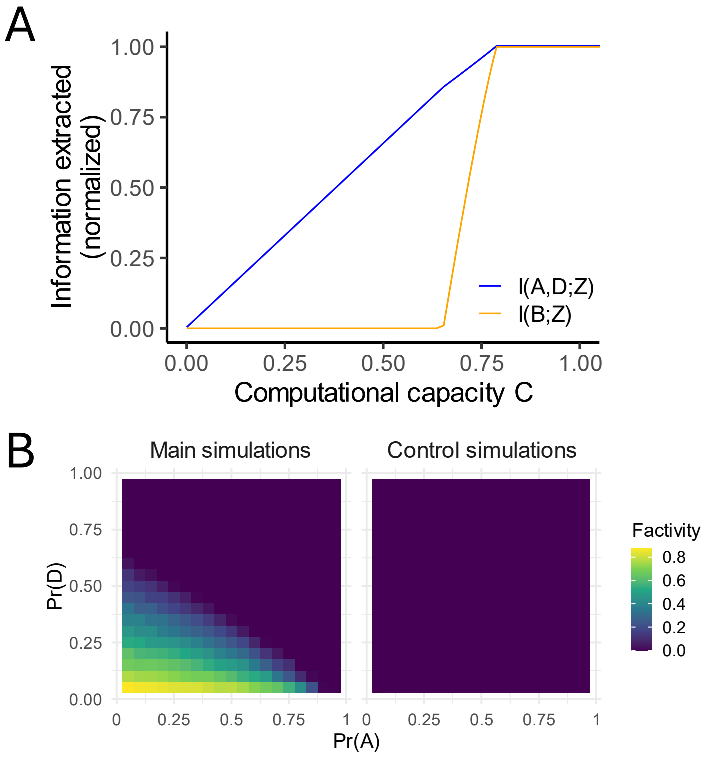

Figure 3: **A:** Amount of information that resource-rational observers extract from knowledge-relevant variables A and D (blue), and belief-relevant variable B (orange), as a function of the observer's computational capacity C, shown here for $N = 3$, $Pr(A) = .2$, $Pr(D) = .05$. Each point on the x-axis corresponds to a different resource-rational observer. Information extracted is normalized such that 1 represents the amount of information extracted by the observer with the largest computational capacity. **B:** Prevalence of factive policies across parameter space, shown for $N = 3$. Factivity is computed as the maximum value of $I(A, D; Z) - I(B; Z)$ across values of $I(X; Z)$, normalized as in A. Intuitively, the brightness of a tile indicates how much higher than the orange line the blue line can get in a plot such as in panel A. In control simulations, the observer does not have a pre-existing representation of S.

theoretic cost of extracting B increases with the number of possible beliefs the agent could have.

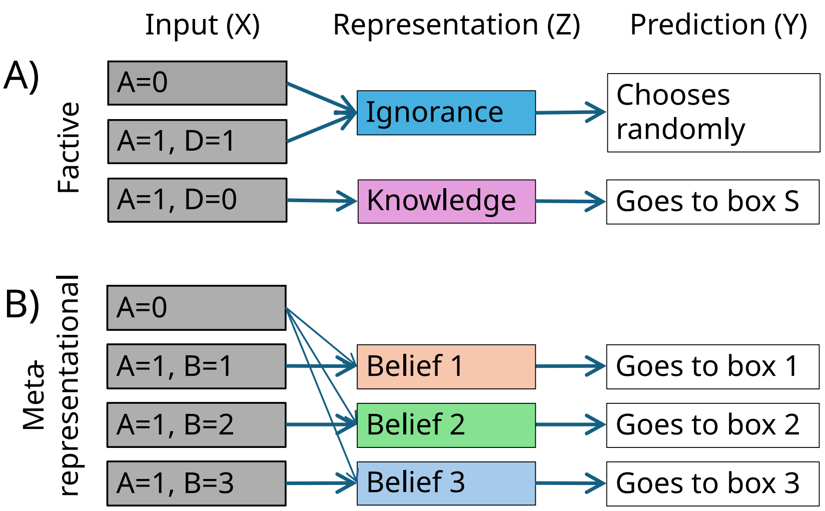

Figure 4: Schematic description of the $\vec{X} \to Z \to \hat{Y}$ mapping in a factive observer (**A**) and a meta-representational observer (**B**), shown for $N = 3$. The meta-representational observer in the state $A = 0$ maintains a uniform probability distribution over the three possible belief states. Note that actual policies are typically more stochastic than these simplified mappings, see Figure S3.

3.1 Experiments

Here we take a closer look at the performance of resource-rational observers by performing 'in-silico' experiments in our mindreading tasks. We also compare these results to existing empirical findings in similar tasks in human and non-human primates. We present results for three observers, a representative each of an 'automatic' policy ($C = 0$), a low-resource- ($C = .5$), and a high-resource observer ($C = 1$). The low-resource observer is of special interest because it is a factive mindreader, as can be seen in Figure 3A. We use a social ecology with $N = 3$ boxes, $Pr(A) = .2$ and $Pr(D) = .05$, and report experiments for other social ecologies in the Supplementary Information.

3.2 Predicting behavior

In our main series of tasks, the observer has to predict which box the agent will reach toward. In **Experiment 1**, the observer must predict the behavior of an agent who knows the location of the item ($A = 1$, $D = 0$, upper-left on Figure 5). We find that all observers correctly predict that the agent will reach for the box containing the item, although this inference is stronger in observers with more cognitive resources. This result mirrors experiments in adults, children, and non-human primates; individuals from these populations can attribute knowledge, but human adults do so more reliably [23, 48, 51–54].

In **Experiment 2**, the observer predicts the behavior of an agent who has a false belief ($A = 1$, $D = 1$, and $b \neq s$, upper-right on Figure 5). Only the high-resource observer correctly predicts that the agent will reach for the item where it last saw it. The low-resource agent is mostly agnostic, maintaining an almost uniform distribution over boxes, with only a slight bias toward the actual location of the item. This pattern again reflects experimental results: human adults can pass false belief tasks while non-human primates usually fail them ([23, 53, 55, 56], but see [57]). Moreover, non-human primates fail false-belief tasks in the same way as the low-resource observer: they find each outcome equally surprising, including seeing the agent go toward a box where the item was never located [58]. Young human children also struggle with false belief tasks, although they fail in a slightly different way than the low-resource observer, because they predict that the agent will look for the item at its actual location [59].

In **Experiment 3**, the agent is ignorant ($A = 0$, lower-right in Figure 5). The high-resource and low-resource observers correctly predict that the agent might go toward any box. Similarly, experimental data show that human adults, young human children and non-human primates predict that an agent that ignores the location of a desirable item might not reach toward it [52, 60, 61].

Experiment 4 has the structure of a 'Gettier case' in epistemology [43]. Outside of the agent's awareness, the item is removed from its original box but then put back in exactly the same box; as a result the agent has an *accidentally true belief* ($A = 1$, $D = 1$, $s = b$, upper-right on Figure 5). While the high-resource observer succeeds at the task, the low-resource observer expects that the agent might look at any location. The pattern of results for the low-resource observer is similar to that of non-human primates [52, 54], who also fail to represent an agent's belief if that belief is true only by luck. Similar patterns have been observed in human children ([52, 62, 63], but see [64]).

3.3 Learning about the world

In our second series of tasks, we look at whether observers can solve the 'inverse' problem of predicting the location of an item from the agent's behavior. This series of tasks is motivated by a recent proposal that the proper evolutionary function of factive mindreading is social learning, and not social prediction [23]. We suggest that although

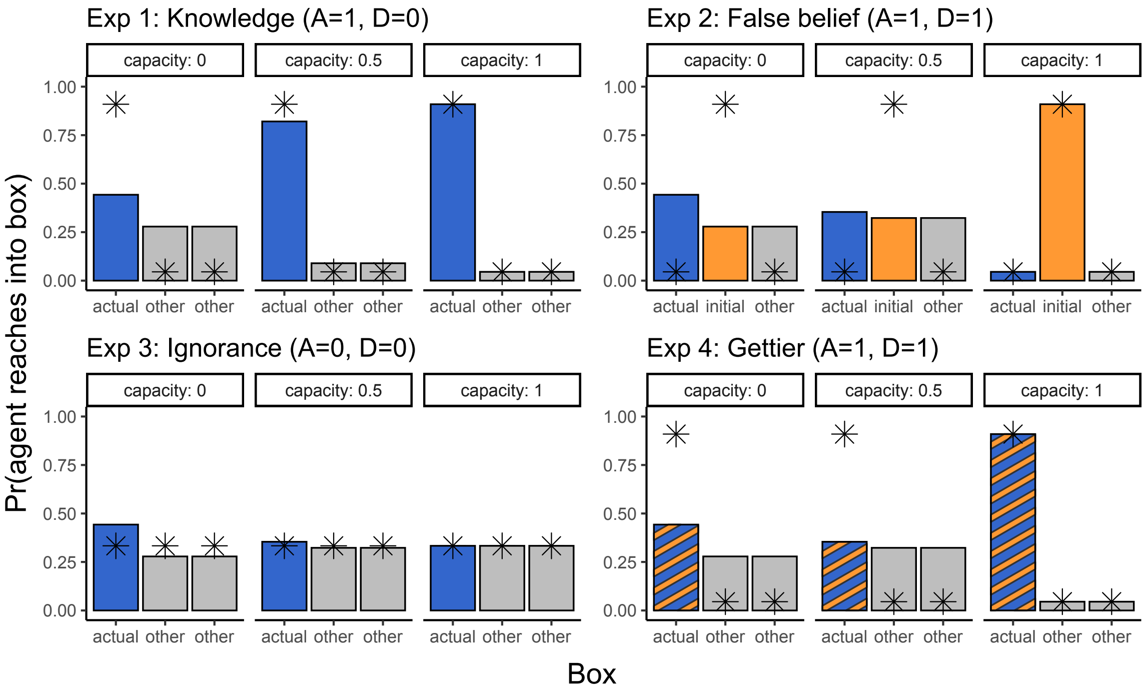

Figure 5: Predictions made in our experiments by resource-rational observers with different computational capacity. 'Actual': actual location s of the item; 'Initial': initial location b of the item. Stars represent the ideal non resource-limited policy. Parameters used were $N = 3$, $Pr(A) = .2$, $Pr(D) = .05$.

factive mindreading is indeed helpful for social learning, this does not necessarily mean that it evolved primarily for that purpose. Specifically, we demonstrate that good social learning performance can also emerge simply as a byproduct of optimizing for social prediction. We show that the representations optimized for our *first* task (predicting behavior from the state of the world), can also be co-opted for predicting the state of the world from observation of another individual's behavior. Formally, in this new task the observer must predict the true location of the item (S) in a situation where they know the agent's choice (Y), whether the agent had perceptual access (A), whether the agent was deceived (D), but don't know either the initial (B) or current location of the item (S).

Using Bayes' rule, a rational observer without cognitive limitations can predict S as (see SI):

$$Pr(S|Y, A, D) = Pr(Y|S, A, D) \tag{4}$$

Since resource-limited observers do not have access to the true generative model, they have to substitute $q(\hat{Y}|S, A, D)$:

$$Pr(S|Y, A, D) = q(\hat{Y}|S, A, D) \tag{5}$$

where $q(\hat{Y}|S, A, D) = \sum_z q(\hat{Y}|S, Z)q(Z|A, D)$. We assume that the encoder $q(Z|A, D)$ and decoder $q(\hat{Y}|S, Z)$ are optimized for our main task of predicting the agent's behavior, i.e. we take them off-the-shelf from the policies derived in our first series of experiments. As above, we look at an automatic policy, a low- and a high-resource observer ($C \in \{0, .5, 1\}$), and use the same parameters ($N = 3$, $Pr(A) = .2$, $Pr(D) = .05$). See Supplementary Information for other points in parameter space.

In **Experiment 5**, the agent knows the location of the item ($A = 1$, $D = 0$), upper-left Figure S5). We find that all three observers correctly infer that the item is located in the box that the agent is reaching toward, although this inference is stronger in observers with more cognitive resources. Similarly, non-human primates infer that a rewarding item is in a box if another agent approaches that box [56].

In **Experiments 6 and 7**, the low- and high-resource observers correctly judge that the agent's behavior is not diagnostic about the item's location if the agent has information that is not up-to-date ($A = 1$, $D = 1$, upper right on Figure S5), or if the agent is ignorant ($A = 0$, lower-left). In contrast, in all conditions the automatic observer predicts that the item is in the box that the agent is reaching toward.

In sum, the low-resource observer (a factive observer) performs close to the ideal policy in all three experiments of our social learning task, despite the fact that its representations were optimized solely for the separate task of predicting behavior. These experiments also illustrate that factive observers in our model can behave adaptively in situations of *egocentric ignorance*, where another agent knows something that they don't [18].

Note that our analysis does not preclude the possibility that social learning has also been an evolutionary pressure for the evolution of factive mindreading. We explore this proposal in more detail in the Supplementary Information.

3.4 High-resource observers flexibly switch between knowledge and belief representation.

High-resource observers in our simulations successfully pass false belief and Gettier tasks. This finding might indicate that high-resource observers implement a fully meta-representational strategy: they encode the content B of the other agent's belief whenever that agent has perceptual access ($A = 1$), see Figure 4 lower panel. Alternatively, high-resource observers might use a flexible strategy: they only encode the content of the agent's belief when the agent was deceived, i.e. when $A = 1$, $D = 1$, and track the agent's knowledge otherwise. This strategy allows an observer to perfectly predict behavior, while potentially saving cognitive resources.

To assess which strategy more closely describes high-resource observers in our simulations, we computed the amount of information that an observer extracts about variables A, D and B, relative to the maximum possible information that can be extracted about that variable (its Shannon entropy). For high values of $Pr(D)$, high-resource observers approximate a fully-metarepresentational strategy, extracting a high portion of the available information about B. In contrast, for low values of $Pr(D)$, high-resource observers approximate a fully-flexible strategy: they represent knowledge by default, and only encode the content of an agent's belief when this agent has a false or accidentally true belief (Figure S4).

3.5 Control simulations

We claim that factive mindreading can be adaptive because observers are already representing the state of the world S, and so can use this information at no extra cost for predicting others' behavior. In the Supplementary Information we report a complementary set of simulations where we abandon this assumption, and find that factive mindreading does not emerge when observers must pay the additional cost of representing S for mindreading-specific purposes—showing that this assumption is indeed essential to our results.

4 Discussion

Social prediction can be computationally expensive. How do organisms with limited cognitive resources efficiently navigate their social environments? Our resource-rational analysis uncovers a key insight: mindreaders can exploit the substantial overlap between their own world model and those of other agents in the same environment to preserve cognitive resources. They can track the facts to which other agents have epistemic access (what they 'know'), instead of explicitly representing the content of other agents' world models (what they 'believe'). This *factive* mindreading strategy was automatically discovered by our resource-rational analysis as the optimal strategy for observers with low cognitive resources, across many different social ecologies.

Our analysis is consistent with a wide range of empirical findings about human and non-human mindreading. As other researchers have argued, young children and non-human primates appear to often or even predominantly engage in factive mindreading [1, 18, 19, 23]. While researchers still debate whether young human children and non-human primates can represent false beliefs [2, 57, 59, 65–67], knowledge representation is much easier than belief representation for these populations, as well as for human adults with autism or under cognitive load [see 23, for extensive evidence]. Similarly, factive observers in our simulations easily pass knowledge attribution tasks, but fail belief attribution tasks.

Crucially, the difficulty of false-belief tasks appears to come from the difficulty of representing belief per se, rather than the difficulty of representing *false* belief in particular. Chimpanzees and rhesus macaques fail to predict the behavior of agents with an *accidentally true belief*, in experiments that implement the equivalent of 'Gettier' cases [43]. In these experiments, an agent sees an item placed into a box, but an experimenter later takes the item out of the box before putting it back into the same box (outside the agent's awareness, but witnessed by the participant). Participants fail to predict that the agent will go towards the item's location [52, 54]. In this paradigm, the agent has a true belief, so a difficulty with representing false beliefs cannot explain non-human primates' failure to predict the agent's behavior [19, 23]. In contrast, the observed pattern is a signature of factive mindreading. Factive mindreaders do not assign knowledge to an agent by checking whether the agent's belief matches reality: instead they attribute knowledge on the basis of reliable cues such as perceptual access, and they reverse this attribution when the situation changes outside of the agent's awareness.

We reproduce these Gettier cases experiments in our simulations and find that factive observers behave just like non-human primates, in that they have no expectation about where the agent will look for the item. Factive observers also exhibit other specific patterns found in non-human primate mindreading: for example they find each outcome equally surprising in a false-belief task, including seeing the agent go toward a box where the item was never located [58].

Our factive observers have a less direct fit to the behavior of young human children: although young children do tend to have an easier time representing knowledge than belief, they tend to fail false belief tasks differently than our model. Specifically, they tend to predict that the agent will look for the item in its actual location [59]. And while some children appear to fail to represent accidentally true belief [62], this failure might be explained by pragmatic confounds rather than competence issues [64]. Overall, our model rests at a high level of abstraction, and a full account of the successes and failures of mindreading (across different ages and species) should also integrate other factors at different levels of analysis [e.g. 11, 68].

Our resource-rational framework is consistent with the advantage that human adults have over children and non-human primates in their capacity for belief representation. The former have higher information-processing capacity than the latter [69], and can allocate more computational resources to mindreading, enabling the costlier approach of belief representation.

At the same time, even human adults have computational limitations, and we expect that they take advantage of the efficiency of factive mindreading at least some of the time. High-resource observers in our model often use a partially factive strategy: they represent knowledge by default, and only encode the content of the agent's belief when a knowledge representation would not allow them to accurately predict behavior (cases of false or accidentally true belief). Human adults might use a similar strategy. Supporting this hypothesis, people can engage in mental state inference in contexts like conversation that require quick and spontaneous mindreading [18], even though they find it difficult to compute beliefs in these same contexts [70]. Similarly, if human adults used a purely meta-representational strategy, we would expect them to judge whether an agent knows something by first computing the agent's belief, and then assessing whether this belief matches reality. Available evidence argues against this proposal: knowledge attribution in human adults is a process distinct from belief attribution [71]. When people are asked what an agent knows, they respond either as fast or *faster* than when they are asked what the agent believes [71, 72]. Similarly, neural activity in knowledge attribution tasks does not exhibit the signatures of inhibitory processing found in belief attribution tasks [71, 73].

We kept our model as simple as possible to make it easy to interpret. Future work could scale up our approach to more complex tasks using a combination of information theory and deep learning [36, 74]. Besides inferring epistemic states, other components of mindreading like goal inference can be approached with a resource-rational lens [75–77]. Future research should also explore tasks beyond social prediction, given the role of mindreading in influencing the behavior of other agents [78, 79].

While we focus on factive mindreading because it has been the subject of previous empirical research, the general principle we identify here is much broader. On our account, mindreaders save computational resources by representing some parts of their own world model as being shared by another agent. In the setting we use here these parts of the world model are facts about the world, but in principle they can be other things, such as concepts. For instance Alice might assume that Bob's concept of APPLE is the same as her own, instead of creating a meta-representation of Bob's concept of APPLE. We suspect that much of social cognition relies on such strategies, and that meta-representation is the exception rather than the norm.

5 Methods

The constrained optimization problem in the main text is equivalent [44] to the problem of minimizing the following functional, where β is a Lagrange multiplier:

$$F_{\beta}[q(z|\vec{x})] = I(\vec{X}; Z) - \beta I(Z; Y|S) \tag{6}$$

There is not a one-to-one mapping between values of C (upper bound on information extracted from $\vec{X}$) and β, so we compute optimal policies for various values of β and find the policy with a value of $I(\vec{X}; Z)$ closest to C.

To compute the optimal encoder for a given value of β, we use a variant of the Blahut-Arimoto algorithm [44, 80, 81] given by [46], in which we iterate the following update equations until convergence:

$$Pr(z|\vec{x}) \propto Pr(z) \exp(-\beta \Sigma_s Pr(\vec{x}|s) D_{\text{KL}}[Pr(y|\vec{x}, s)||Pr(y|z, s)]) \tag{7}$$

$$Pr(z) = \Sigma_{\vec{x}} Pr(z|\vec{x}) Pr(\vec{x}) \tag{8}$$

$$Pr(y|z, s) = \Sigma_{\vec{x}} Pr(y|z, s, \vec{x}) Pr(\vec{x}|z, s) \tag{9}$$

where D_{KL} is the Kullback-Leibler divergence, with:

$$D_{\text{KL}}[Pr(y|\vec{x}, s)||Pr(y|z, s)] = \Sigma_y Pr(y|\vec{x}, s) \log \frac{Pr(y|\vec{x}, s)}{Pr(y|z, s)} \tag{10}$$

To mitigate the fact that the information bottleneck method does not guarantee convergence to a global optimum, we use the method of reverse deterministic annealing [82], see Supplementary Information.

References

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

[2] Baillargeon R, Scott RM, Bian L. Psychological reasoning in infancy. Annual review of psychology. 2016;67(1):159-86.

[3] Gergely G, Csibra G. Teleological reasoning in infancy: The naıve theory of rational action. Trends in cognitive sciences. 2003;7(7):287-92.

[4] Saxe R, Kanwisher N. People thinking about thinking people. The role of the temporo-parietal junction in ”theory of mind”. NeuroImage. 2003;19(4):1835-42.

[5] Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. Development of the social brain from age three to twelve years. Nature communications. 2018;9(1):1027.

[6] 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(4):0064.

[7] Quillien T, German TC. A simple definition of 'intentionally'. Cognition. 2021;214:104806.

[8] 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.

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

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

[11] Rakoczy H. Foundations of theory of mind and its development in early childhood. Nature Reviews Psychology. 2022;1(4):223-35.

[12] Lieder F, Griffiths TL. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behavioral and brain sciences. 2020;43:e1.

[13] Craik KJW. The nature of explanation. vol. 445. Cambridge University Press; 1943.

[14] Conant RC, Ross Ashby W. Every good regulator of a system must be a model of that system. International journal of systems science. 1970;1(2):89-97.

[15] Richens J, Everitt T. Robust agents learn causal world models. arXiv preprint arXiv:240210877. 2024.

[16] Leslie AM. Pretense and representation: The origins of "theory of mind.". Psychological review. 1987;94(4):412.

[17] Sperber D. Metarepresentations: A multidisciplinary perspective. Oxford University Press; 2000.

[18] Westra E, Nagel J. Mindreading in conversation. Cognition. 2021;210:104618.

[19] Nagel J. Factive and nonfactive mental state attribution. Mind & Language. 2017;32(5):525-44.

[20] Williamson T. Knowledge and its Limits. Oxford University Press; 2002.

[21] Phillips J, Norby A. Factive theory of mind. Mind & Language. 2021;36(1):3-26.

[22] Levy Y. The priority of intentional action: From developmental to conceptual priority. The Philosophical Quarterly. 2024:pqae023.

[23] Phillips J, Buckwalter W, Cushman F, Friedman O, Martin A, Turri J, et al. Knowledge before belief. Behavioral and Brain Sciences. 2021;44:e140.

[24] Gershman SJ, Horvitz EJ, Tenenbaum JB. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science. 2015;349(6245):273-8.

[25] Lewis RL, Howes A, Singh S. Computational rationality: Linking mechanism and behavior through bounded utility maximization. Topics in cognitive science. 2014;6(2):279-311.

[26] Icard T. Bayes, bounds, and rational analysis. Philosophy of Science. 2018;85(1):79-101.

[27] Sims CR. Rate–distortion theory and human perception. Cognition. 2016;152:181-98.

[28] Wei XX, Stocker AA. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. Nature neuroscience. 2015;18(10):1509-17.

[29] Sims CR, Jacobs RA, Knill DC. An ideal observer analysis of visual working memory. Psychological review. 2012;119(4):807.

[30] Gershman SJ. The rational analysis of memory. In: Oxford handbook of human memory. Oxford University Press Oxford, UK; 2021. .

[31] Futrell R. Information-theoretic principles in incremental language production. Proceedings of the National Academy of Sciences. 2023;120(39):e2220593120.

[32] Zaslavsky N, Hu J, Levy RP. A rate-distortion view of human pragmatic reasoning. arXiv preprint arXiv:200506641. 2020.

[33] Taylor-Davies M, Lucas CG. Balancing utility and cognitive cost in social representation. arXiv preprint arXiv:231004852. 2023.

[34] Sims CA. Implications of rational inattention. Journal of monetary Economics. 2003;50(3):665-90.

[35] Polanía R, Woodford M, Ruff CC. Efficient coding of subjective value. Nature neuroscience. 2019;22(1):134-42.

[36] Binz M, Schulz E. Modeling human exploration through resource-rational reinforcement learning. Advances in neural information processing systems. 2022;35:31755-68.

[37] Lai L, Gershman SJ. Human decision making balances reward maximization and policy compression. PLOS Computational Biology. 2024 04;20:1-32. Available from: https://doi.org/10.1371/journal.pcbi.1012057.

[38] Ortega PA, Braun DA. Thermodynamics as a theory of decision-making with information-processing costs. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2013;469(2153):20120683. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2012.0683.

[39] Arumugam D, Ho MK, Goodman ND, Van Roy B. Bayesian Reinforcement Learning With Limited Cognitive Load. Open Mind. 2024 04;8:395-438. Available from: https://doi.org/10.1162/opmi_a_00132.

[40] Cheyette SJ, Wu S, Piantadosi ST. Limited information-processing capacity in vision explains number psychophysics. Psychological Review. 2024.

[41] Icard T, Goodman ND. A Resource-Rational Approach to the Causal Frame Problem. In: Proceedings of the cognitive science society; 2015. .

[42] Kinney DB, Lombrozo T. Building Compressed Causal Models of the World. Cognitive Psychology. 2023.

[43] Gettier E. Is justified true belief knowledge? Analysis. 1963.

[44] Tishby N, Pereira FC, Bialek W. The information bottleneck method. arXiv preprint physics/0004057. 1999.

[45] Berger T. Rate-distortion theory. Wiley Encyclopedia of Telecommunications. 2003.

[46] Gondek D, Hofmann T. Conditional information bottleneck clustering. In: 3rd ieee international conference on data mining, workshop on clustering large data sets; 2003. p. 36-42.

[47] Zaslavsky N, Kemp C, Regier T, Tishby N. Efficient compression in color naming and its evolution. Proceedings of the National Academy of Sciences. 2018;115(31):7937-42.

[48] Horschler DJ, Berke MD, Santos LR, Jara-Ettinger J. Differences Between Human and Non-Human Primate Theory of Mind: Evidence from Computational Modeling. bioRxiv. 2023:2023-08.

[49] Todd PM, Gigerenzer G. Ecological rationality: Intelligence in the world. OUP USA; 2012.

[50] Simon HA. A behavioral model of rational choice. The quarterly journal of economics. 1955:99-118.

[51] Pillow BH. Early understanding of perception as a source of knowledge. Journal of experimental child psychology. 1989;47(1):116-29.

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

[53] Marticorena DC, Ruiz AM, Mukerji C, Goddu A, Santos LR. Monkeys represent others' knowledge but not their beliefs. Developmental science. 2011;14(6):1406-16.

[54] 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.

[55] Martin A, Santos LR. The origins of belief representation: Monkeys fail to automatically represent others' beliefs. Cognition. 2014;130(3):300-8.

[56] Krachun C, Carpenter M, Call J, Tomasello M. A competitive nonverbal false belief task for children and apes. Developmental science. 2009;12(4):521-35.

[57] 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.

[58] Royka AL, Horschler DJ, Bargmann W, Santos L. Probing Nonhuman Primate Errors on False Belief Tasks to Explore the Evolutionary Roots of Theory of Mind. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 46; 2024. .

[59] Wimmer H, Perner J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition. 1983;13(1):103-28.

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

[61] Townrow L, Krupenye C. Bonobos point more for ignorant than knowledgeable social partners. Proceedings of the National Academy of Sciences. 2025;122(6).

[62] Fabricius WV, Boyer TW, Weimer AA, Carroll K. True or false: Do 5-year-olds understand belief? Developmental Psychology. 2010;46(6):1402.

[63] Fabricius WV, Gonzales CR, Pesch A, Weimer AA, Pugliese J, Carroll K, et al. Perceptual access reasoning (PAR) in developing a representational theory of mind. Monographs of the Society for Research in Child Development. 2021;86(3):7-154.

[64] Oktay-Gür N, Rakoczy H. Children's difficulty with true belief tasks: Competence deficit or performance problem? Cognition. 2017;166:28-41.

[65] Butterfill SA, Apperly IA. How to construct a minimal theory of mind. Mind & Language. 2013;28(5):606-37.

[66] Heyes C. Submentalizing: I am not really reading your mind. Perspectives on Psychological Science. 2014;9(2):131-43.

[67] Burge T. Do infants and nonhuman animals attribute mental states? Psychological Review. 2018;125(3):409.

[68] Leslie AM, Friedman O, German TP. Core mechanisms in 'theory of mind'. Trends in cognitive sciences. 2004;8(12):528-33.

[69] Cantlon JF, Piantadosi ST. Uniquely human intelligence arose from expanded information capacity. Nature Reviews Psychology. 2024;3(4):275-93.

[70] Keysar B, Lin S, Barr DJ. Limits on theory of mind use in adults. Cognition. 2003;89(1):25-41.

[71] Bricker AM. The neural and cognitive mechanisms of knowledge attribution: An EEG study. Cognition. 2020;203:104412.

[72] Phillips J, Knobe J, Strickland B, Armary P, Cushman F. Evidence for evaluations of knowledge prior to belief. In: Proceedings of the cognitive science society; 2018. .

[73] Gonzalez B, Armary P, Dungan J, Strickland B, Knobe J, Cushman F, et al. Knowledge without belief. 2025. Available from: https://osf.io/preprints/psyarxiv/ht65f_v2.

[74] Alemi AA, Fischer I, Dillon JV, Murphy K. Deep variational information bottleneck. arXiv preprint arXiv:161200410. 2016.

[75] Blokpoel M, Kwisthout J, van der Weide TP, Wareham T, van Rooij I. A computational-level explanation of the speed of goal inference. Journal of Mathematical Psychology. 2013;57(3-4):117-33.

[76] Chandra K, Chen T, Li TM, Ragan-Kelley J, Tenenbaum J. Inferring the future by imagining the past. Advances in Neural Information Processing Systems. 2023;36:21196-216.

[77] Zhi-Xuan T, Kang G, Mansinghka V, Tenenbaum JB. Infinite Ends from Finite Samples: Open-Ended Goal Inference as Top-Down Bayesian Filtering of Bottom-Up Proposals. Proceedings of the Annual Meeting of the Cognitive Science Society. 2024 Jul;46(46).

[78] Ho MK, Saxe R, Cushman F. Planning with theory of mind. Trends in Cognitive Sciences. 2022;26(11):959-71.

[79] Sell A, Sznycer D, Al-Shawaf L, Lim J, Krauss A, Feldman A, et al. The grammar of anger: Mapping the computational architecture of a recalibrational emotion. Cognition. 2017;168:110-28.

[80] Blahut R. Computation of channel capacity and rate-distortion functions. IEEE transactions on Information Theory. 1972;18(4):460-73.

[81] Arimoto S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Transactions on Information Theory. 1972;18(1):14-20.

[82] Zaslavsky N, Tishby N. Deterministic annealing and the evolution of optimal information bottleneck representations. 2019.