Contents lists available at ScienceDirect

Cognition

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

Full Length Article

Naïve epistemics: A theory of rational and error-prone mental state reasoning

Branden J. Bio[a,b,*], Sangeet Khemlani[a,*]

[a] *Navy Center for Applied Research in Artificial Intelligence, U.S. Naval Research Laboratory, USA*
[b] *National Research Council Research Associate, National Academies of Sciences, Engineering, and Medicine, USA*

ARTICLE INFO

Keywords:
Omniscience errors
Mental states
Epistemic reasoning
Knowledge
Belief
Mental models

ABSTRACT

Effective communication depends on reasoning about what others know and believe, and failures in executive functioning can disrupt the way adults reason about mental states. Studies reveal that failures in interpreting premises, simulating possibilities, and formulating conclusions can all yield systematic errors in reasoning – but no account exists of the specific sorts of error people produce when these failures occur in the context of mental state reasoning. We developed such a theory to account for both rational and error-prone mental state reasoning. The theory makes three proposals: first, people build representations of possibilities, and tag those representations, to distinguish knowledge from belief; second, they update, inspect, and consolidate representations of possibilities to engage in mental state reasoning; and third, they can integrate semantic contents into their representations of belief states by constructing or else blocking the construction of alternative possibilities. We tested the theory by examining the patterns of conclusions reasoners produced using a novel sentence construction interface or else through free response. These generative tasks permitted analyses of participants' tendency to draw sensible epistemic conclusions as well as their systematic errors, and they corroborate the central tenets of the theory.

1. Introduction

Effective communication depends on mental state reasoning – for instance, speakers must track what a listener already knows to avoid repeating information (Frank & Goodman, 2012; Jara-Ettinger & Rubio-Fernandez, 2021). Social intelligence requires people to maintain belief states over longer time horizons: to plan a surprise party for a friend, you must communicate information to guests so the surprisee remains unaware. Many occupational contexts – the law, national security, and public relations, to name a few – demand that their professionals engage capably in mental state reaosning; a public relations officer who routinely exposes sensitive information is unlikely to keep their job. Indeed, studies on jurors suggest that they take into account mental state information even when instructed otherwise (e.g., Margoni & Brown, 2023).

The ability to represent belief states develops piecemeal throughout childhood (Wellman, 2018; Woo, Chisholm, & Spelke, 2024) and children make many errors during the process, e.g., they mistakenly think that other people have access to their knowledge, as revealed by the "Sally-Anne" task (Baron-Cohen, Leslie, & Frith, 1985), and they have trouble representing others' perspectives (Surtees, Butterfill, & Apperly, 2012). These deficits may come about as a consequence of limitations in other processes such as executive functioning (Kouklari, Thompson, Monks, & Tsermentseli, 2017) and inhibitory control (Austin, Groppe, & Elsner, 2014; Sabbagh, Xu, Carlson, Moses, & Lee, 2006). Even mature reasoners have difficulty separating their beliefs from others. Birch and Bloom (2004, 2007) document a "curse of knowledge" bias in which a person's knowledge of the consequence of some event compromises their ability to reason about other people's beliefs (see also Diamond & Kirkham, 2005; Royzman, Cassidy, & Baron, 2003; Shah & LaForest, 2022; Tullis & Feder, 2023). Reasoners compartmentalize beliefs systematically (Apperly, Samson, & Humphreys, 2009; Keysar, Lin, & Barr, 2003), and recent studies reveal neural substrates where such compartmentalization can occur (Bio, Guterstam, Pinsk, Wilterson, and Graziano, 2022; Thornton, Weaverdyck, & Tamir, 2019). The brain appears to recruit machinery for representing one's own mental states to represent those of others (e.g., Bio, Webb, and Graziano, 2018; Kovács, Téglás, & Endress, 2010).

* Corresponding author at: Navy Center for Applied Research in Artificial Intelligence, US Naval Research Laboratory, 4555 Overlook Dr. SW, Washington, DC 20375, USA.

E-mail addresses: brandenbio@gmail.com (B.J. Bio), sangeet.khemlani.civ@us.navy.mil (S. Khemlani).

https://doi.org/10.1016/j.cognition.2025.106328
Received 16 October 2024; Received in revised form 10 September 2025; Accepted 15 September 2025

Many cognitive, communicative, and computational accounts of reasoning assume that mature adults, in aggregate and even when burdened by significant stress, can reason about mental states in an optimal way: they have few systematic difficulties withholding information, tracking others' beliefs, or keeping secrets. Hence, efficient communication can be modeled as interactions between rational, pragmatic speakers who arrive at probabilistic assumptions about each other's mental states (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Frank & Goodman, 2012). One recent account argues that people rapidly infer communicative intent when they hear mental state language (Jara-Ettinger & Rubio-Fernandez, 2021). The authors examined how a speaker and a listener jointly communicate to resolve ambiguous relational descriptions of objects in a scene, and they implemented their theory as a form of probabilistic inference over both the speaker's set of beliefs about the environment as well as their intentions of what to say. The model successfully explained participants' abilities to infer the object referenced by a speaker in a visual display of multiple objects across a wide variety of conditions. The account was intended to explain how human inferences can mirror those made by rational agents that communicate effectively with one another, but not as a model of the processes by which they do so, or as a comprehensive theory of mental state reasoning.

Like other forms of conscious reasoning, mental state reasoning processes are subject to the capacity limitations of executive functioning and working memory (Baddeley, Hitch, & Allen, 2021; Bouchacourt & Buschman, 2019; Logie, Camos, & Cowan, 2020; Peloquin, 2021; Unsworth & Robison, 2020) – significant cognitive load can disrupt adults' abilities for tracking belief states (Schneider, Lam, Bayliss, & Dux, 2012). Reasoners under stress may therefore devote significant mental resources to tracking and compartmentalizing mental states, particularly for salient and sensitive matters. For that reason, a robust account of mental state reasoning should investigate the errors adults make when assigning mental states to others.

In what follows, we argue against the assumption that mature adult reasoning is optimal and error-free – indeed, the process of compartmentalizing representations of others' beliefs may routinely introduce errors. To date, no studies have investigated errors specific to the processes that underlie mental state reasoning, namely, the processes by which people interpret what an agent knows or believes, and the processes by which they mentally simulate those beliefs. That may be because mature reasoners learn to monitor and correct routine errors and deliberate about them when necessary. It may also be because there are consequences – social, professional, legal – to improperly tracking what others know and don't know, and those consequences can serve to reinforce capable reasoning about epistemic matters. Yet, if systematic reasoning errors exist, they may provide insights into the processes humans use to use to mentally simulate belief states and which procedures of the reasoning process are prone to fault.

We accordingly propose a novel theory designed to account for both optimal and suboptimal mental state reasoning in adults. The theory assumes that people construct iconic possibilities and maintain them in working memory to reason about epistemic matters by tagging such possibilities with information concerning the mental states of others. It predicts that individuals should often generate false conclusions about others' mental states because of inappropriate assignment of such epistemic tags and describes the mechanisms by which they can assign tags appropriately. It can therefore serve as a foundation for higher level accounts of theory of mind and communicative intent (e.g., Jara-Ettinger & Rubio-Fernandez, 2021). We describe the theory in detail in the next section and evaluate experiments that test its predictions in the subsequent section. We conclude by discussing potential challenges to the theory, its implications, and what error-prone mental state reasoning reveals about how people mentally represent the minds of others.

2. A theory of reasoning about mental states

The theory of epistemic reasoning we describe shares a central assumption with other accounts of human thinking: people reason by mentally simulating possible situations in the world (see Johnson-Laird, 1983 for the first such proposal and Carey, Leahy, Redshaw, & Suddendorf, 2020; Gerstenberg, 2024; Johnson-Laird & Ragni, 2024; Johnson-Laird & Ragni, 2019; Phillips, Morris, & Cushman, 2019 for recent possibility-based theories). One such account – the mental model theory of reasoning – has explored how people reason with possibilities across many different domains, including causal (Goldvarg & Johnson-Laird, 2001), spatial (Ragni & Knauff, 2013), deontic (Bucciarelli & Johnson-Laird, 2005), and sentential reasoning (Khemlani, Byrne, & Johnson-Laird, 2018). The theory argues that when a reasoner constructs a possibility – either through perception, imagination, or language comprehension – they build an idealized, simplified analog of the information available. This tradeoff permits reasoners to draw conclusions rapidly, but it necessitates that they discard information irrelevant to their reasoning goals (Knauff & Johnson-Laird, 2002; see also Bigelow et al., 2023). The theory proposes four complementary ideas about how people represent and reason with possibilities:

i. **Models of possibilities are iconic.** Representations of possibilities – *mental models* – mimic the structure of the real-world situations they represent (Peirce, 1931–1958, Vol. 4). An iconic model of a spatial relation involving three objects, such as *the circle is between the square and the triangle*, consists of three tokens that represent those shapes in an appropriate configuration. Reasoners scan models to make inferences (Ragni & Knauff, 2013). Models can represent both static and dynamic scenarios that unfold in time (Khemlani et al., 2013).

ii. **People represent abstract concepts with symbolic tags.** Many abstract concepts cannot be represented in an iconic way: negations, for instance, do not correspond to specific real-world scenarios, e.g., *the circle isn't next to the square* cannot be represented by any specific spatial configuration or set of configurations. Reasoners represent negations by tagging models with symbols (Khemlani, Orenes, & Johnson-Laird, 2012) and processing those symbols to consider alternative models (Khemlani, Orenes, & Johnson-Laird, 2014; Orenes, Beltrán, & Santamaría, 2014). Likewise, models can be tagged with numerical values to represent numerical premises (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999).

iii. **People tend to draw conclusions from a single initial possibility they construct.** After constructing a single possibility representing a set of premises, reasoners tend not to consider alternative models – doing so imposes a tax on working memory (Johnson-Laird & Khemlani, 2023). Problems that require individuals to consider multiple models are therefore harder than those requiring only one model (see, e.g., Johnson-Laird & Byrne, 1991; Johnson-Laird & Khemlani, 2023; Kelly, Khemlani, & Johnson-Laird, 2020). Reasoners privilege explanations based on one possibility (Korman & Khemlani, 2020) and they coalesce, simplify, and reduce possibilities in other ways that facilitate inference (Johnson-Laird & Ragni, 2024).

iv. **They consider alternative possibilities through serial deliberation.** To make optimal inferences, reasoners need to search for possibilities beyond the initial one they construct. They do so by making modifications to initial models and checking those modifications serially; hence, reasoning difficulty increases with the number of modifications needed to solve a particular problem (Cortes et al., 2021; Ragni, Khemlani, & Johnson-Laird, 2014). Individual differences, pragmatics, and problem contents can affect the tendency to consider multiple possibilities (Johnson-Laird & Byrne, 2002), and computational models of the theory simulate such variation by stipulating the stochastic properties of model construction and search (Khemlani & Johnson-Laird, 2022).

A viable theory of human reasoning should account for what makes some reasoning problems easy and others difficult. The model theory argues that difficulty comes from people's tendency to overlook possibilities. They do so in at least three contexts: during the process of interpreting premises into a semantics that is useful for constructing possibilities, which is biased to consider typical possibilities over atypical ones (Jahn, Knauff, & Johnson-Laird, 2007; Khemlani, 2018; Ragni & Knauff, 2013); during the process of constructing models, which tends to ignore possibilities that could make premises false (Johnson-Laird, 2010; Khemlani & Johnson-Laird, 2017); and when modifying initial models to search for alternatives, which is a recursive, memory-greedy process that can halt before it considers all relevant possibilities (Khemlani & Johnson-Laird, 2022). Shortcuts in any of these processes can reduce demands on working memory, but they lead to systematic patterns – preferences, biases, and errors – that alternative theories cannot explain (Johnson-Laird & Khemlani, 2023; Kelly & Khemlani, 2023; Khemlani & Johnson-Laird, 2017), including errors in processing sentential connectives such as "if" and "or" (Khemlani, 2018).

Some inferences are easy. Consider the conclusion in (1):

1. If Olga is a client, then she's a student.
   Olga is a client.
   Therefore, she's a student.

The inference, known as *modus ponens*, is easy enough that children produce it for abstract contents by 10 years of age, if not much earlier (Markovits & Barrouillet, 2002); nearly all adults accept such arguments (Schroyens, Schaeken, & d'Ydewalle, 2001) and have no difficulty with them (Byrne, Evans and Newstead, 1993; Byrne et al., 2019). Mental models explain why the problem is easy. People represent the first premise by constructing a single model, which we illustrate using the following diagram:

```
client          student                    possibility #1
          ...
```

where the words are tokens that stand in place of the contents of the possibilities. It depicts the possibility in which Olga is a client and also a student. The ellipsis ('…') serves as a placeholder that other possibilities are consistent with the premise. Reasoners will disregard it until they need to consider what follows when Olga is not a client. Models are updated by incrementally combining models of each new premise (Johnson-Laird & Khemlani, 2023; Khemlani, Wasylyshyn, Briggs, & Bello, 2018). The second premise asserts that Olga is a client, which is already represented in the model – and so the conclusion that Olga is a student follows from inspecting the rest of the model (in bold).

Other inferences are more difficult, such as this *modus tollens* inference:

2. If Mia is a client, then she's a student.
   Mia is not a student.
   Therefore, Mia is not a client.

When asked to draw their own conclusions, reasoners often respond erroneously that nothing follows (Johnson-Laird, Byrne, & Schaeken, 1992; Wason & Johnson-Laird, 1972); likewise, they accept the conclusion in (2) only 74% of the time (Schroyens et al., 2001). When they do infer that Mia is not a client, they are slower to do so relative to *modus ponens* problems (Barrouillet, Grosset, & Lecas, 2000). The model theory accounts for these patterns. Reasoners initially construct a single model to represent the first premise, just as above:

```
client          student                    possibility #1
          ...
```

The second premise states that Mia is not a client – and so it eliminates the initial model, which is what causes many participants to conclude that nothing follows. Those who deliberate longer do so by fleshing out their initial model to consider alternative possibilities consistent with the conditional premise:

```
  client          student                   possibility #1
¬ client        ¬ student                   possibility #2
¬ client          student                   possibility #3
```

When combined with the factual premise in (2), the result is a single possibility:

```
¬ client      ¬ student
```

which yields the inference that Mia is not a client (in bold). The extent to which people spontaneously draw such inferences depends on numerous factors, such as their background knowledge of disabling conditions (Cummins, Lubart, Alksnis, and Rist, 1991; Johnson-Laird & Byrne, 2002). Indeed, the model theory predicts that knowledge affects model construction by blocking possibilities and introducing relations (see Khemlani, Wasylyshyn, Briggs, and Bello, 2018). For example, this *modus tollens* inference is easier because of the contents of the premises:

3. If Mia is a client, then pigs will fly.
   Pigs won't fly.
   Therefore, Mia is not a client.

The contents in the *then*-clause stipulate a situation that is false and meant to be disregarded. The theory predicts that people should take these contents into account by *modulating* their interpretation of the conditional so that it represents only the singular possibility in which pigs don't fly – which is also the possibility in which Mia is not a client:

~~client   student~~   possibility #1 (blocked)
¬ client   ¬ pigs fly   possibility #2
~~¬ client   student~~   possibility #3 (blocked)

And results show that contents do indeed modulate how people interpret conditionals; they reveal patterns that corroborate the predictions of the model theory (Quelhas, Johnson-Laird and Juhos, 2010).

Standard systems of logic – those based on the sentential calculus – treat both (1) and (2) above as *valid* (Jeffrey, 1981, p. 1), that is, when the premises are true, the conclusions must be true, too. Logics use the concept of validity to distinguish patterns of reasoning that are truth preserving from those that could potentially introduce falsehoods and contradictions. Yet they do not adjudicate between easy and difficult inferences: they cannot explain why (1) is easier than (2), or why (2) is harder than (3). They also cannot distinguish sensible inferences from "vapid" ones: any set of premises permits an infinity of vapid, but valid, logical inferences (Johnson-Laird, Khemlani, & Goodwin, 2015). Consider this inference:

4. Olga is a student.
   Therefore, Olga is student or she is a programmer.

The inference is of the form:

4′. A.
   Therefore, A or B.

where *A* and *B* stand for any proposition whatsoever. It is valid in any standard system of logic, and yet people reject it (Hinterecker, Knauff, & Johnson-Laird, 2016; Orenes & Johnson-Laird, 2012), because it isn't necessary that Olga is a programmer: you can hypothesize scenarios in which she's not. Reasoners similarly accept certain inferences even though they are invalid in any standard system of "modal" logic, i.e., a logic designed to cope with the concepts of possibility and necessary. Consider this inference:

5. Olga is a student or she is a programmer.
   Therefore, it's possible that she is a programmer.

People accept it at ceiling (Hinterecker, Knauff, & Johnson-Laird, 2016) – but it is invalid in all normal modal logics, because if it's in fact impossible that Olga is a programmer then the premise could be true while the conclusion is false. The model theory, in contrast, treats the disjunction in (5) as inherently modal, i.e., the disjunction refers to a set of possibilities, one of which is that Olga is a programmer. Hence, it predicts the widespread acceptance of the conclusion.

We can apply similar analyses, not to inferences about facts or possibilities, but also epistemic matters of knowledge and belief. Consider the following argument:

6. Olga knows that if there's an ace in the deck, then there's a queen in the deck.
   Olga knows that there's an ace in the deck.

It is similar to the *modus ponens* inference in (1), but it uses the verb *know* in both premises to refer to a set of mental states, namely the knowledge Olga maintains about the deck. It is sensible to draw these conclusions from (6):

7a. Therefore, Olga knows that there's a queen in the deck.
 b. Therefore, there's a queen in the deck.

Olga can easily infer the presence of a queen from the information in (6), which permits the conclusion in (7a). And since the verb *knows* is a "factive" verb that presupposes the truth of its complement – we return to the concept of factivity below – it presupposes that there is, in fact, a queen in the deck. In contrast, these are unacceptable conclusions:

8a. Therefore, Mia knows that there's a queen in the deck.
 b. Therefore, Olga knows that there's a queen or a king in the deck.

The conclusion in (8a) concerns Mia's mental states, not Olga's, and nothing in (6) provides any information about what Mia knows. So (8a) does not follow. (8b) is more subtle: intuitively, (6) does not provide any information about whether there's a king in the deck – it is possible that there is, but not necessary.

One way to characterize reasoning about epistemic states is to appeal to an epistemic logic (Bolander, 2018; van de Pol, van Rooij, & Szymanik, 2018; van Ditchmarsch & Labuschagne, 2007), which is a system of logic designed to formalize the concepts of knowledge and belief (Fagin, Halpern, Moses, & Vardi, 1995; von Wright, 1951; Hintikka, 1962). Many epistemic logics treat (6) as valid and (7a) as invalid (Luper, 2016; see also Leuenberger & Smith, 2021). They implement some variation of a "closure axiom", which is designed to capture the intuition that agents can competently draw deductive conclusions for themselves. Cohen (2002, p. 312) stipulates a closure axiom as follows: "If *S knows P* and *S knows P entails Q*, then *S knows Q*", where *S*, *P*, and *Q* stand in place of any proposition. If you replace *S* with "Olga", *P* with "ace in the deck", and *Q* with "queen in the deck", then the axiom has a structure similar to (6) and (7a) above, and so it sensibly treats (7a) as valid. But the axiom treats (8b) as valid, too, because *there's a queen or a king in the deck* is a valid entailment of *there's a queen in the deck* on any orthodox logic.

The moral of the story is that systems of logic do not adjudicate on what makes an inference easy, difficult, meaningful, or sensible. They cannot serve as the basis of any theory of human reasoning, including mental state reasoning (Johnson-Laird, Byrne and Khemlani, 2024). The model theory, however, explains all these assessments by assuming that people base their inferences on the construction and manipulation of iconic possibilities. A conclusion is acceptable if it is necessary, i.e., it holds in all possibilities consistent with the premises. It is possible if it holds in at least one such possibility; and it is probable if it holds in most of them (Bell & Johnson-Laird, 1998). The model theory therefore serves as the foundation of a new theory of mental state reasoning.

2.1. Reasoning about mental states: A model-based theory

Psychological theories of reasoning have yet to explain the mental processes by which people make or reject mental state inferences like those above, and so no account explains why people would infer (7a) and (7b) but reject (8a) or (8b). We therefore introduce a model-based account of how people encode, represent, and reason with mental states. The theory we develop posits three novel principles, which we summarize briefly:

- **The principle of mental state tags:** reasoners encode the mental states of an agent by constructing models of non-mental state content and then tagging those models with information about the agent who holds them.
- **The principle of consolidation:** people reason by consolidating models tagged by an agent's mental states; they update and scan consolidated models to draw conclusions. If they can draw one or more conclusions from a set of consolidated models, then by default they ascribe those conclusions to the corresponding agent.
- **The principle of alternatives:** reasoners distinguish knowledge from belief based on how they consider alternative models: beliefs permit reasoners to entertain all possible contingencies, while knowledge constrains the construction of possibilities.

We describe each principle in turn.

2.1.1. Mental state tags

Despite earlier arguments that all beliefs are encoded in a centralized memory store (Quine & Ullian, 1978), recent theorists treat beliefs as though they are stored in fragmented, contextually organized memory clusters (see Bendaña & Mandelbaum, 2021; Elga & Rayo, 2022; Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000; Zhao, Richie, & Bhatia, 2022) including those that maintain counterfactual beliefs, imagined future possibilities, and states of desire and intention (Byrne, 2005, 2017; De Brigard, 2023; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Harner & Khemlani, 2022). This "fragmentation" hypothesis helps to explain how Olga's beliefs could be kept separate from Mia's: they may be represented in different belief clusters. The following principle imports this hypothesis and describes how models of belief can be encoded in memory:

> **The principle of mental state tags.** When reasoners interpret a mental state verb, such as *know*, *believe*, *think*, and *assume*, to ascribe a state of mind to an individual, they construct a model corresponding to the contents of the mental state along with a *tag* that represents the agent who maintains it. People reason about others' mental states by updating and inspecting similarly tagged models. Reasoners distinguish factive tags, which allow inferences from mental states to facts, and non-factive tags, which do not.

The tag principle accordingly treats the construction of the model as well as the application of a tag as two separate and distinct processes, each of which are subject to error. For instance, reasoners may tend to combine information and associate a single tag with a single model. As we describe below, such behavior can lead to systematic error. The principle states that reasoners distinguish between factive and non-factive tags. Throughout the remainder of the paper, we denote tags using black backgrounded text associated with a possibility, as in:

```
ace                                    [Olga]
```

This diagram represents the possibility of an ace in the hand paired with a mental state tag, where the brackets denote that Olga holds this information as knowledge. And this diagram:

```
          queen                  (Mia)
```

uses parentheses to represent Mia's belief that the queen is in the hand. The specific conventions for separating between factive and non-factive states are immaterial (see, e.g., Khemlani, 2021 for an alternative). What matters instead is that models can be used to preserve and track information about belief states.

We illustrate the tag principle by considering how reasoners may interpret (6) above:

Olga knows that if there's an ace in the deck, then there's a queen in the deck.

```
ace            queen                   [Olga]
         . . .
```

Olga knows there's an ace in the deck.

```
ace                                    [Olga]
```

These models are akin to those presented for (1) above, but we use black backgrounded text to represent the mental state tag associated with each possibility. The model of the conditional and the model of the fact both have the same tag, which specifies that they're both pieces of knowledge held by Olga. And the brackets of the tag denote that it is factive, and so it presupposes the truth of the tagged possibility. The product of the two models is therefore:

```
ace            **queen**               [Olga]
```

and inspection of this model yields the conclusion in (7a) and bolded above: *Olga knows that there's a queen in the deck*. The presence of the tag biases reasoners to consider Olga's mental states over the facts they might infer, but since there exists only one model that represents a factive tag, reasoners can use it to infer facts about the world.

Consider what the tag principle predicts of these premises, which are a variation of (6) above:

9. Toni knows that if there's an ace in the deck, then there's a queen in the deck.
   Olga knows that there's an ace in the deck.

Reasoners should construct models of the conditional premise and associate a tag with Toni's knowledge:

```
ace            queen                   [Toni]
         . . .
```

and they should do the same with Olga's knowledge:

```
ace                                    [Olga]
```

but they should keep the two models separate. Since both of the premises are factive, those who deliberate further may construct a third model of the facts at hand, which combines all presupposed information:

```
          ace            queen
```

But constructing this model should place a burden on working memory, because it requires reasoners to track three separate states of affair: what Toni knows, what Olga knows, and what their combined knowledge implies. Reasoners who maintain this information can make nuanced inferences from (9):

10a. It's possible that Olga doesn't know that there's a queen in the hand.

b. It's possible that Toni may not know that there's an ace in the hand.

Those who do not may not realize that they can draw a definitive inference about what's in the hand. They may instead focus on one agent's knowledge over another. As a consequence, reasoners should be more likely to conclude that there's a queen in the deck from (6) compared to (9).

A more striking consequence of the tag principle is that, because model construction and tagging are two separate processes, tags can be inappropriately applied, particularly in cases when models are updated with new information. Consider how you might respond to this problem:

11. Toni knows that if there's an ace in the deck, then there's a queen in the deck.
    There's an ace in the deck.
    What, if anything, follows?

Nothing in the premises asserts that Toni knows about any of the cards in the deck – but the first premise presupposes the truth of the conditional. Hence, reasoners should draw the following conclusion:

12. Therefore, there's a queen in the deck.

which necessarily holds. To make this inference, they need to construct a model of Toni's knowledge as well as a separate model which integrates that knowledge with the facts at hand. Because this process may be burdensome, many reasoners may opt for a shortcut by constructing a model of the first premise above:

```
ace            queen                   [Toni]
         . . .
```

And then *updating* that single model with the information from the second premise. In effect, they have erroneously treated the facts of the matter as pertaining to Toni's knowledge, and so may make this conclusion instead:

13. Toni knows that there's a queen in the deck.

We refer to the conclusion in (13) as an *omniscience* error, because it reveals a failure to appropriately compartmentalize mental states, and it leads to undue omniscience on Toni's part (see Appendix A for a logical analysis of such errors). Experiments 1a-e and 2 show that omniscience errors were common conclusions that participants drew from problems such as (11).

Epistemic concepts are fundamentally recursive; the following statement is sensible and meaningful:

14. Carol believes that Sam believes that he is brave.

The statement concerns Sam's and Carol's beliefs, and Carol could possess at least two separate beliefs about the situation: her own belief about whether Sam is brave or not, and her belief about Sam's opinion on the matter. Indeed, studies on false beliefs, theory of mind, and second-order false beliefs take the recursive nature of mental states for granted (Arslan, Hohenberger, & Verbrugge, 2017; Bianco et al., 2021; Miller, 2009). The interactions between Sally and Anne in a task often used to test for such thinking (Baron-Cohen et al., 1985; Wimmer & Perner, 1983) can be summarized as follows:

15. Sally places a marble in a basket.
    So, Sally believes the marble is in the basket.
    She leaves.
    Anne moves the marble from the basket into a box.
    So, Anne knows the marble is in the box.

The description in (15) concerns several states of belief: Anne's knowledge about the position of the marble, and your belief about Sally's (false) belief. You can also consider additional belief states, such as Anne's belief about Sally's belief. The only boundary on a reasoner's recursive structure of beliefs is their limited working memory. Because epistemic concepts permit recursion, epistemic verbs – both factives and non-factives – demand a recursive syntax (Miller & Johnson-Laird, 1976): they require a complement clause, which itself can make use of an epistemic verb, as in (14) above. Factive verbs are therefore absent in languages without recursion, such as Pirahã (Everett, 2012).

The model theory's iconicity principle (see above) combined with its tag principle can help explain how people engage in meta-epistemic reasoning, that is, reasoning about one agent's mental states concerning another agent's mental states. They do so by using mental state information to construct a recursively structured model. The following diagram depicts a mental model of (14) above:

```
┌──────────────────────────┐
│  Sam is brave   (Sam)    │   (Carol)
└──────────────────────────┘
```

Such that the box represents a model of the beliefs that Sam holds about himself, tagged accordingly. Tags can be applied to the contents of the model, as well as the model as a whole to represent another agent's beliefs about Sam's beliefs. An emergent consequence of this representation is that Carol's beliefs about Sam's bravery are held separate from her beliefs about Sam's beliefs. That is, she could concur or else disagree, but her opinion would demand the construction of another model. The theory accordingly predicts that reasoning about (14) should be easier than (16) below:

16. Carol believes that Sam believes that he is brave.
    But Carol believes that Sam is cowardly.

since (14) concerns one model while (16) concerns two. The prediction is difficult to test because the difference is confounded by the number of premises in the two examples. So, we examined a testable consequence of the theory instead. Consider these variations of (14):

17a. Carol knows that Sam knows that he is brave.
   b. Carol believes that Sam knows that he is brave.
   c. Carol knows that Sam believes that he is brave.

The model theory argues that they differ from one another in the models they yield. The model of (17a) is:

```
┌──────────────────────────┐
│  Sam is brave   [Sam]    │   [Carol]
└──────────────────────────┘
```

It uses factive tags to represent what Sam knows and what Carol knows about what Sam knows. How does this representation differ from the model of (16) above? An immediate consequence is that reasoners should be more likely to conclude that Sam is brave from (17a) than (16). The complement of "Carol knows that" is, "Sam knows that he is brave" – and this complement presupposes the truth of his bravery. The same is true for (17b), i.e., it should support conclusions about Sam's bravery more often than (17c), because (17b) yields a model in which a factive tag applies to a model of Sam's bravery, whereas in (17c), the factive tag applies to a model of Sam's mental state. Experiment 3 tested and corroborated these predictions.

2.1.2. Consolidation

Consider again (9) above. One of the reasons this problem may be more difficult for reasoners is the need to keep Toni and Olga's mental states separate: Toni knows something Olga doesn't and vice versa. The corollary is that reasoners should draw mental state inferences when they are able to consolidate information appropriately. Consider this variation:

9′. Toni knows that if there's an ace in the deck, then there's a queen in the deck.
    Olga knows that there's an ace in the deck and conveys that information to Toni.

What follows from (9′)? Clearly, Toni can infer that there's a queen in the deck but Olga cannot. The model theory explains this disparity as a consequence of consolidating tagged information: once we know that Olga conveyed her knowledge to Toni, we assume Toni has it too – and we apply the appropriate tag. Hence, we can construct two separate models, one that captures the conditional information:

```
   ace            queen                 [Toni]
           . . .
```

and a second that captures Olga and Toni's shared knowledge:

```
 ace                              [Olga] [Toni]
```

The model theory posits a consolidation process in which reasoners integrate and coalesce similarly tagged models. Hence, Toni's knowledge becomes:

```
   ace            queen                 [Toni]
```

and so she is able to make the necessary inference that there's a queen in the deck.

The following principle describes this consolidation process:

> **The principle of consolidation.** People reason about an agent's mental states by consolidating models tagged and associated with that agent. To reason deductively, they may consolidate only those models that have factive tags; to reason inductively, they may consolidate factive tags with non-factive tags. They update and scan consolidated models to draw conclusions. If they cannot construct a model – such as when some information is inconsistent – then reasoners refrain from making inferences concerning that information. Otherwise, if they can draw one or more conclusions from a set of consolidated models, then by default they ascribe those conclusions to the corresponding agent.

The process of consolidation may be protracted and deliberative: reasoners may choose to reject relevant information, integrate background information not pertaining to an agent's beliefs, or else integrate information that an agent could easily discover to simulate some of the agent's potential inferences. We illustrate the process with several examples.

In (9′) above, one agent conveys information to another, and we can assume that the other agent took the information as knowledge. But consider this additional variation:

9″. Toni knows that if there's an ace in the deck, then there's a queen in the deck.
    Olga suspects that there's an ace in the deck and conveys that belief to Toni.

What should Toni conclude from (9″)? The process of consolidation is such that consolidating knowledge with a belief results in a belief. If Toni accepts Olga's information at face value, they both should believe, but not know for sure, that there's a queen in the deck. More generally, if tags are mismatched, they should cause people to reason inductively by tagging models and preserving non-factive tags over factive ones. Consider this similar problem:

18. Ruhi knows that she'll attend Ohio State or Penn State.
    Ruhi believes that she won't attend Penn State.

Since Ruhi only believes she won't attend Penn State, we cannot conclude that she'll attend Ohio State for certain, and we certainly cannot conclude that she knows as much. But we can conclude: Ruhi believes that she'll attend Ohio State. The principle of consolidation shows how this inference is feasible. The first premise in (18) yields these models:

```
Ohio State                              [Ruhi]
                  Penn State            [Ruhi]
```

and the second premise yields this one:

```
¬ Penn State        (Ruhi)
```

Since the tags mismatch, their consolidation yields this model:

```
Ohio State                              (Ruhi)
```

which yields the predicted inference. Experiment 4 tested and corroborate this prediction for analogous conditional premises.

What happens when an individual attempts to convince another of something false or impossible? Consider this example:

19. Pia knows that if there's a jack in the deck, then there's a king in the deck.
    Yuri conveys to Pia that there's a jack but not a king in the deck.

Examples such as these reveal a breakdown in the consolidation process: Pia cannot integrate Yuri's insights into her understanding of the deck. The model theory explains this difficulty as a failure of consolidation, if tags are applied to each of these models:

```
jack         king              [Pia]
       ...
jack       ¬ king              (Yuri) (Pia)
```

to capture Pia's potential belief, then it is impossible to consolidate and combine Yuri's information with Pia's conditional knowledge, i.e., it is impossible to build an integrated model. Reasoners may adopt a number of strategies to cope with this inconsistency, e.g., they may attempt to explain Yuri's actions, or else infer Pia's rejection or disbelief in Yuri's information. Model consolidation – and the lack thereof – can provide insights into how people may infer that Yuri conveyed false information. Suppose that the second premise in (19) is replaced with the following:

Yuri convinces Pia that there's a jack but not a king in the deck.

In this case, we may interpret *convince* in model-theoretic terms: it refers to a situation in which one agent successfully causes another to adopt a belief, i.e., it refers to the application of a non-factive tag to a model. Yet, since reasoners cannot consolidate Pia's old knowledge with what Yuri convinced her of, they should reject that knowledge in some way, e.g., by inferring that Pia no longer believes that if there's a jack in the deck, there's a king in the deck. If they maintained the presupposition that the conditional is true – even though Pia no longer believes it – they may further infer that Yuri convinced Pia of something false.

2.1.3. *Alternative models*

Consider how you might answer (20) below:

20. Ari believes, but doesn't know, that the meeting is on Wednesday or else Thursday.
    Is it possible that the meeting is on Monday?

Ari's belief could be mistaken, so the answer to (20) is yes. The model theory explains how reasoners comprehend this information. They may initially construct a model of Ari's disjunctive belief:

```
Wednesday                                (Ari)
                 Thursday                (Ari)
```

by building models of both possibilities and adding appropriate epistemic tags, which use parentheses to denote that they are non-factive. Non-factive tags permit reasoners to infer additional contingencies, i.e., they can infer possibilities corresponding to both the contents of the mental state as well as their negations. So, while Ari believes the meeting is in two days, in fact there are four possibilities to consider:

```
  Wednesday
                 Thursday
¬ Wednesday
               ¬ Thursday
```

These possibilities can be further elaborated by using background knowledge to conclude that the meeting could be held on any day of the week. Ari's belief does not make any presuppositions about what is true in the real world – yet the information is meaningful, because it can be used to detect conflicts in other agents' belief. Consider Jen's belief in (21):

21. Jen believes, but doesn't know, that the meeting is on Tuesday or else Friday.

An initial model of Jen's beliefs is accordingly:

```
Tuesday                                  (Jen)
                 Friday                  (Jen)
```

Of course, Jen can be just as mistaken as Ari about the day of the meeting – and yet reasoners can make inferences, not just about whether beliefs adhere to the facts or not, but whether they are consistent relative to another person's set of beliefs. Indeed, they can track disagreements and conflicts between different agents and between a single agent and the facts of the matter. Hence, supposing that (20) and (21) are both true, these conclusions necessarily follow regardless of who is right and who is wrong:

22a. Ari's and Jen's beliefs about the day of the meeting cannot both be true at the same time.
b. If Ari's beliefs are right, then Jen's are wrong. And if Jen's are right, then Ari's are wrong.

Models explain how people detect inconsistencies (Johnson-Laird et al., 2000; Johnson-Laird, Girotto, & Legrenzi, 2004): they attempt to build a single model of both of the agents' beliefs. If they can construct such a model, the two agents hold consistent beliefs; if not, their beliefs are in conflict (see Harner & Khemlani, 2022). Reasoners may further deliberate to construct an explanatory model that resolves the conflict by serving as an alternative model of the scenario (Khemlani & Johnson-Laird, 2011), and they may appeal to epistemic concepts when constructing such a model (Kelly & Khemlani, 2023, Table 1).

The present theory proposes that reasoners distinguish beliefs and knowledge on the basis of the alternative models they can generate. It assumes the following principle:

The principle of alternatives. Reasoners rapidly construct initial models of the contents of mental states – but they can consider

> alternative models as well. Non-factive tags permit them to consider all possible contingencies concerning belief contents and their negations, as well as any elaboration of those contents or negations. In contrast, factive tags – which are used to refer to an individual's knowledge – block reasoners from constructing possibilities that contradict the contents of the knowledge. The more mental state possibilities that reasoners have to construct, the more difficult it is to reason.

To illustrate the principle, consider the initial models for (20) above. The principle states that reasoners can consider all possible contingencies compatible with Ari's beliefs combined with all the contingencies compatible with their negations. One way to negate Ari's beliefs is to consider those possibilities where the meeting is held on multiple days. So, the full set of possibilities should include these:

```
   Monday    ¬ Tuesday    ¬ Wednesday    ...   (Ari)
 ¬ Monday      Tuesday    ¬ Wednesday    ...   (Ari)
 ¬ Monday    ¬ Tuesday      Wednesday    ...   (Ari)
   ...         ...          ...
   Monday      Tuesday    ¬ Wednesday    ...   (Ari)
   Monday    ¬ Tuesday      Wednesday    ...   (Ari)
```

and many others – a total of $7^2 = 49$ possibilities altogether. They are too many for anybody to consider individually. In contrast, consider the relevant possibilities when the verb in (20) is factive instead:

20′. Ari knows that the meeting is on Wednesday or else Tuesday.

In this case, the initial models are:

```
   Wednesday                       [Ari]
                Thursday           [Ari]
```

Reasoners can flesh these possibilities out to explicitly represent what is false, i.e., that a meeting on Wednesday implies that there's no meeting on Monday, Tuesday, and so on. But they should not consider possibilities apart from these two options.

As these examples make clear, an immediate consequence of the principle is that knowledge should be easier to reason about than belief (see Phillips et al., 2019): in typical discourse circumstances, people should find problems concerning the verb *know* easier than those concerning the verb *believe* (Nazlidou et al., 2018). Experiments 1a and 1b revealed this pattern – they found fewer omniscience errors for factive verbs than non-factive verbs. The principle further explains why children master factives earlier than non-factives – factives require the construction and maintenance of fewer models – and why they initially misinterpret verbs such as *believe* to be factive (Abbeduto & Rosenberg, 1985; see also Shatz, Wellman, & Silber, 1983). And it is consistent with neurobehavioral data from discourse comprehension tasks, which show that comprehenders resist incorporating information that violates a factive verb's presuppositions, and that they are more likely to incorporate that information for non-factives (Shetreet, Alexander, Romoli, Chierchia, & Kuperberg, 2019). Shetreet and colleagues conclude, in line with the present account, that "the presuppositions triggered by factive verbs are encoded and maintained within the comprehender's discourse model."

Table 1
The three principles of the model theory of epistemic reasoning, along with central predictions they make, and the experiments that test them.

Principle	Prediction	Experiment
The principle of mental state tags	People should spontaneously produce omniscience errors	Experiments 1a-e and 2
	People should comprehend and reason about meta-epistemic relations, i.e., one agent's thoughts about another agent's thoughts, in a manner that reflects tag structure	Experiment 3
The principle of consolidation	People should generate conclusions about an agent's mental state more often than conclusions from presuppositions	Experiment 4
The principle of alternatives	People's reasoning about online and offline knowledge should interact to make certain inferences necessary when they would otherwise not be	Experiment 5

We accordingly consider a more subtle prediction of the theory: reasoners should distinguish between *online* knowledge, that is, the facts and presuppositions they acquire from comprehending mental state language and discourse, and *offline* knowledge, which concerns the commonsense facts and information encoded in declarative, episodic, and semantic memory (Kumar, 2021; Renoult, Irish, Moscovitch, & Rugg, 2019; Squire, 2004) and recalled as needed. Consider (23a) and (23b):

23a. Pterodactyls are not dinosaurs (they are pterosaurs).
b. Hildegarde is aware that pterodactyls are pterosaurs.

The statement in (23a) expresses offline knowledge, i.e., a fact about how scientists classify pterodactyls. In contrast, (23b) concerns *online* knowledge, i.e., an expression about the knowledge an individual possesses, in this case using the factive verb *aware*. Both online and offline knowledge operate similarly and in accordance with the principle of alternatives in two ways: first, they prevent the consideration of alternative possibilities, such as the possibility that pterodactyls are dinosaurs or that Hildegarde believes as such (Johnson-Laird & Byrne, 2002). They can also introduce relational dependencies (Juhos, Quelhas, & Johnson-Laird, 2012). Consider these examples which introduce spatiotemporal relations:

24a. He studied for the test but failed it.
b. Wang discovered pterosaur eggs in the Gobi Desert.

because individuals possess offline knowledge about what "studying" means, they can infer from (24a) that he studied for the test before he failed it. And because the meaning of *discover* refers to a transition from a state of ignorance to knowledge, they can infer from the online knowledge in (24b) that at some point in time, Wang didn't know the eggs' location and at a later point in time, he did.

The principle of alternatives predicts that offline and online knowledge interact with one another to modulate reasoning. We illustrate the interaction in what follows. Consider this problem:

25. Devon knows that if it's an animal, then it's hidden.
Devon knows that it's not an animal.
Does it follow that it's not hidden?

If you focus only on the underlined text in (25), the inference maps onto the following pattern: *If A then B; Not A; Therefore, not B.* This pattern – called *denial of the antecedent* – is invalid in logic, though reasoners endorse such inferences, and often do so in error (Johnson-Laird et al., 1992); that is, the conclusion below:

26. If it's an animal, then it's hidden.
It's not an animal.
Therefore, it's not hidden.

does not follow, because even though it's not an animal, it could be something else that's hidden, such as a hidden electronic device. Hence, (26) is compatible with the following models:

```
¬ animal       hidden
¬ animal     ¬ hidden
```

The same argument holds for (25) above: Devon's knowledge of these facts rules out some possibilities, but it allows for the two above. And (27) below allows for even more possibilities:

27. Devon believes, but doesn't know, that if it's an animal, then it's hidden.
    Devon believes, but doesn't know, that it's not an animal.
    Does it follow that it's not hidden?

because it is compatible with those possibilities in which the conditional is false and those possibilities in which an animal is present. Contrast (25–27) with this example:

28. Devon knows that if it's an animal, then it's a wolf.
    Devon knows that it's not an animal.
    Does it follow that it's not a wolf?

The example uses "it's a wolf" in place of "it's hidden" in (25). This change matters because reasoners have background knowledge that wolves are animals, which blocks the construction of any possibility in which something is a wolf but not an animal. Hence, (28) is compatible with only this possibility:

~~¬ animal wolf~~
¬ animal ¬ wolf

and the correct answer to it is "yes". Experiment 5 tested and corroborated this prediction.

We next describe each of the studies that tested the principles of the theory.

3. Empirical tests of the theory

Table 1 summarizes the theory's three principles and the predictions they make. We evaluated the theory in a series of studies designed to test each principle. This section presents each study and discusses its results in light of what it reveals about the theory.

3.1. Testing the principle of mental state tags

The principle of mental state tags predicts that reasoners should produce errors of omniscience, in which they inappropriately infer from (10) some conclusion about a mental state. One challenge in testing for the presence of such errors is to present participants with a neutral way in which to describe their conclusions. If we were to ask a participant to assess an erroneous conclusion, they may accept it even if they would never spontaneously produce it. To obviate this concern, we developed a novel sentence completion task; Appendix B provides an example and overview of the interface. We ran a series of studies – Experiments 1a-e and Experiment 2 – that revealed that participants spontaneously produced omniscience errors. The tag principle predicts that they should do so for both factive and non-factive verbs, and so each of these studies manipulated verb factivity. The results corroborate the principle and show how people fail to compartmentalize the mental states of the agents they consider from their own deductive inferences.

The principle of mental state tags likewise predicts that people should engage in meta-epistemic reasoning to distinguish between sentences such as (14) and (17). Experiment 3 tested and discovered differences between such sentences in line with the tag principle's predictions.

3.1.1. Experiments 1a-e

A pilot study (available online at https://osf.io/s24ak/ and described in Bio & Khemlani, 2023a) presented participants with a description of visitors at a wildlife park who have heard some animal sound during their visit. They responded to problems such as:

29. Riley believes that if it's 6:30 pm, then [the sound] is a frog.
    It's 6:30 pm.
    What, if anything, follows?

The vast majority of participants' responses were omniscience errors akin to: "Riley believes that the sound is a frog." This form of omniscience error is particularly egregious for two reasons: first, Riley's conditional belief could be mistaken, and second, the premises do not state that Riley has access to the current time. Yet, some participants may assume she does, e.g., because she may wear a watch or carry a cellphone, and so the pilot may have encouraged participants to produce such responses. Experiments 1a-d therefore sought to test for and eliminate any such pragmatic considerations. They described a scenario in which students are looking for clues in a scavenger hunt across the various rooms of their school. For half of the trials, participants responded to problems such as this:

30. Taylor is in the study.
    The library is locked and inaccessible.
    Taylor knows that if a globe is in the library, then the password is pear.
    A globe is in the library.
    What, if anything, follows?

The correct answer is that "the password is 'pear'" – yet the theory predicts that participants should respond with an omniscience error of the form: "Taylor knows that the password is 'pear'". They should do so despite the pragmatic clues in (30) designed to prevent such errors: Taylor is in a different room than the object described in the *if*-clause, and the relevant room is inaccessible. We accordingly refer to problems such as (30) as *inaccessible* problems, because the agent in the problem has no access to the evidence needed to make the inference that the password is a pear.

The balance of the problems were *accessible* problems, in which participants could reasonably infer that the agent has access to the evidence:

31. Ari is in the library.
    The library is open and accessible.
    Ari knows that if a globe is in the library, then the password is pear.
    A globe is in the library.
    What, if anything, follows?

This problem is identical to (30) in its description of mental states, and yet participants may infer that Ari knows the password because he is in the library and can see the globe. If participants integrate pragmatic and spatial cues into their mental state inferences, then the inaccessible problems should eliminate all omniscience errors. The accessible problems remove these physical barriers to knowledge. The two conditions should therefore yield stark differences in mental state reasoning – but, if participants show a tendency to produce omniscience errors for inaccessible problems, then those errors are likely robust – and egregious. Experiments 1a-e showed that participants produced omniscience errors often for both sorts of problem. The design of each study ruled out various counterarguments for the pattern, which we describe in detail below.

3.1.1.1. Method. Participants. Participants across Experiments 1a-e volunteered through Amazon Mechanical Turk (see Paolacci, Chandler and Ipeirotis, 2010, for a review). They were paid at a rate of $15 USD per hour for these and all subsequent experiments were reported. There were:

- 63 participants in Experiment 1a (mean age = 37.16 years; 26 females, 34 males, 3 prefer not to answer); 9 failed attention checks and were excluded from analysis, yielding data for 54 participants
- 62 participants in Experiment 1b (mean age = 41.87 years; 29 females, 33 males); 19 failed attention checks, yielding 43 participants
- 58 participants in Experiment 1c (mean age = 39.05 years; 26 females, 31 males, 1 prefer not to disclose); 15 failed attention checks, yielding 43 participants
- 65 participants in Experiment 1d (mean age = 41.72 years; 28 females, 36 males, 1 prefer not to disclose); 20 failed attention checks, yielding a total of 45 participants
- 66 participants in Experiment 1e (mean age = 42.98 years; 31 females, 35 males); 11 failed attention checks, yielding a total of 55 participants

Design, procedure, and materials. Participants in Experiment 1a carried out 16 test problems and 2 attention check problems. The experiment manipulated an agent's ability to access information, and it made use of semantic content designed to eliminate errors. The materials concerned a scavenger hunt scenario in which agents explored a building to uncover clues and find passwords. Information about the password depended on an object's presence at a particular location, e.g., "… if a globe is in the library, then the password is pear." Half the problems in Experiment 1a (accessible problems) described agents who were in the same room as the object, as in (31) above; the other half (inaccessible problems) described agents in a different room from the object. Likewise, the experiment manipulated the mental state verb in the premises, such that half the problems used *know*, as in (23–24) above, and the other half used *believe*. Both (30) and (31) depict an epistemic analog of a *modus ponens* problem structure, i.e., one in which the truth of the *if*-clause is asserted. To test whether omniscience errors generalize beyond such problems, half of the problems in Experiment 1a used an epistemic modus ponens premise structure ($\mathrm{MP_E}$) and the other half used an epistemic modus tollens ($\mathrm{MT_E}$) structure, which negates the *then*-clause:

32. Sammy is in the planetarium.
    The library is locked and inaccessible.
    Sammy knows that if a jar is in the library, then the password is pineapple.
    The password is not pineapple.

Hence, Experiment 1a reflected a 2 × 2 × 2 repeated-measures design that manipulated the epistemic verb (*believe* vs. *know*), the problem structure ($\mathrm{MP_E}$ vs. $\mathrm{MT_E}$), and the agent's access to information. The names of the agents in the problem, the names of the rooms, and the passwords were assigned randomly from pools of materials such that no participant saw the same combination of problems and materials across the study. The study randomized the order of the problems and the order of options in the sentence-construction interface. Experimenters can explicitly probe people's inferences by presenting premises and then asking open-ended questions, such as "What, if anything follows?" or "What do you think happened?" or other such questions. They must then code participants' natural responses. The difficulty with adapting such tasks to study mental state reasoning is that participants may not spontaneously describe the inferences they make about mental states, and their natural responses may be difficult to interpret. Evaluative, forced-choice tasks provide possible conclusions for participants to endorse or reject, but in doing so, they may introduce biases and coax participants to consider conclusions that they might not generate spontaneously. To address these limitations, we developed a hybrid, quasi-generative methodology to study mental state reasoning and employed it in the experiments we report. Participants received open-ended prompts and interacted with an interface (see Fig. 1) that permitted them to respond by clicking one or more buttons that corresponded to the various words needed to construct a full sentence. Each time they clicked a button, the corresponding word was added to the end of the sentence, and the button was removed from the options available. Participants could click a "reset" button in case they made an error. They could also click a button corresponding to "nothing follows". The approach allowed participants to consider all the various pieces of information in a given scenario, both relevant and irrelevant; and it minimized indirect effects of response options. The choice of words from which participants built their responses focused on the inferences most relevant to the problems in Experiments 1a-d. But, by presenting participants with sentence fragments instead of entire sentences, the study minimized any bias to respond with descriptions of mental states. This method makes the probability of producing any coherent answer by chance incredibly small, because it allows participants to select any subset of nine possible sentence fragments to construct a sentence – hence, the probability of generating any sentence by chance alone rounds to zero (1 in $\sum_{i=1}^{9} \frac{9!}{(9-i)!} = 804{,}969$). The quasi-generative nature of the task likewise permitted efficient coding of omniscience errors.

The experimental instructions trained participants on sample practice trials that familiarized them with how to build their responses. Instructions showed how the same list of options could be used to make many different kinds of responses to the same problem, i.e., they explicitly encouraged participants to consider multiple response strategies. This sentence construction methodology was also flexible enough to allow for an attention verification within the same general problem design. Attention check trials were nearly identical to the problems in the study with the exception that participants were told to create the nonsensical sentence, "Believes that knows that nothing follows" rather than providing their own response. This provided a seamless transition between experimental problems and attention checks to verify participants' focus on the task.

Once the experiment displayed the premises, there was a 1 s delay before the sentence construction interface appeared. It presented a list of clickable buttons, which allowed participants to create complete sentences by selecting words from those provided. They received no feedback about the sentence they constructed, and they were permitted to construct a sentence or start over as they pleased until they were satisfied with their conclusion. The designs, materials, and procedures for Experiments 1b-e matched those of Experiment 1a, except where we outline here. *Experiment 1b)* To rule out the concern that participants constructed responses to match the surface structure of premises, i.e., by selecting the same words that were used in the problem, Experiment 1b replaced the factive verb *know* with the factive verb *understand* in the conditional premise (e.g., "Sammy understands that if a jar…") and likewise, it replaced the non-factive verb *believe* with the non-factive verb *think*. The response options in the sentence construction interface, however, used the verbs *know* and *believe*, prohibiting any tendency to select matching verbs when constructing a sentence. The experiment was similar to Experiment 1a in every other way. *Experiment 1c)* Experiment 1c was identical to Experiment 1a in every way, except that it added the words "You discover that…" (a second-person factive stipulation) to the last premise to yield a statement such as, "You discover that a globe is in the study." This change sought to create a pragmatic cue that separates the global facts of the matter from the beliefs of any agent described in the premises. *Experiment 1d)* Experiment 1d was identical to Experiment 1a but for two changes: first, it altered the last premise in an attempt to prevent omniscience errors entirely, such that the last premise stated, e.g., "You know that a globe is in the study and *Ari does not*" (italicized here for emphasis). This explicit stipulation was designed to prevent participants from inferring that Ari knows the password. Second, Experiment 1d did not vary participants' access to information, i.e., it depicted all problems in the inaccessible condition. Experiment 1e) Experiment 1e was identical to Experiment 1d in every way – it presented only inaccessible problems and it explicitly denied the agent's knowledge – but it further varied the prompt that participants saw. On half the trials, the prompt referred to what the participant could deduce, e.g., "What, if anything, *can you conclude*?" (emphasis added). On the other half, the prompt referred to

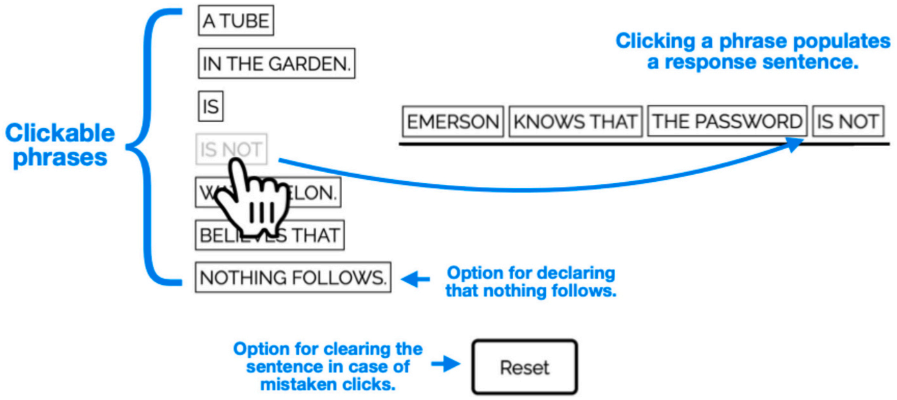
A TUBE
IN THE GARDEN.
IS
IS NOT
W LON.
BELI S THAT
NOTHING FOLLOWS.
Clickable phrases
Clicking a phrase populates a response sentence.
EMERSON
KNOWS THAT
THE PASSWORD
IS NOT
Option for declaring that nothing follows.
Option for clearing the sentence in case of mistaken clicks.
Reset

Fig. 1. Participants received a set of premises such as those in (31, main text) and responded to open-ended questions such as, "What, if anything, follows?" A set of buttons (shown on the left) corresponded to phrases that participants used to construct a response sentences (shown on the right). One button in the interface permitted participants to declare that nothing followed, and another permitted them to clear the response sentence and start the task over in case of mistakes.

what the agent could deduce, e.g., "What, if anything, *can Ari conclude*?" (ditto). This latter version of the prompt, combined with denial of the agent's knowledge, was designed to eliminate omniscience errors.

Open science. Experiment 1a and subsequent experiments were preregistered through OSF (https://osf.io/k873v). Experimental code, materials, coding rubrics, data, and analyses scripts for this study and all subsequent ones can be found at this same link.

Coding rubric. For Experiments 1a-e, we coded responses that contained the agent and either of the epistemic verbs as erroneous, e.g., "Ash knows that…", because no information in the problem structure permitted such conclusions. These responses attribute a mental state to an agent even though the trial stipulated only information about the state of the world, i.e., information that the agent may not have access to. Reasoners could make other sorts of errors, too: they could, e.g., fail to generate a valid inference and erroneously conclude that nothing follows – an error of omission). They could respond in a way that was logically invalid regardless of any consideration of mental states (a logical error). And they could produce either nonsensical or else accurate responses. We coded every response for each of these outcomes. The full coding rubric is available on OSF (https://osf.io/83ja6/).

3.1.1.2. Results. Participants attributed epistemic states to agents even when it was not appropriate to do so: they committed omniscience errors in Experiments 1a-e. Table 2 shows the percentages of such errors, along with other kinds of responses, for each experiment. We describe the results of each study separately.

Experiment 1a. Participants made omniscience errors on 44% of trials in Experiment 1a: 43 out of 54 participants produced at least one such error (binomial test, $p < .0001$ assuming a conservative chance probability of 5%). They yielded such errors 54% of the time for problems that described an agent who had access to information relevant to their mental state compared to 33% of the time when the agent did not have such access (Wilcoxon test, $z = 3.94$, $p < .001$, Cliff's $\delta = 0.29$). Participants were therefore sensitive to the accessibility of the information. Participants yielded omniscience errors more often than chance (i.e., ~0%) for both inaccessible and accessible problems (Wilcoxon tests, $z\text{s} > 5.57$, $p\text{s} < 0.001$, Cliff's $\delta\text{s} > 0.60$). They generated omniscience errors more often for *believe* than for *know* (47% vs. 41%, Wilcoxon test, $z = 2.55$, $p < .001$, Cliff's $\delta = 0.10$). The structure of the problems did not affect their tendency to err: they produced errors 46% of the time for MP_{E} problems and 41% of the time for MT_{E} problems (Wilcoxon test, $z = 0.96$, $p = .34$, Cliff's $\delta = 0.07$).

Each possible two-way interaction was reliable or close to it (Wilcoxon tests, $z\text{s} > 1.89$, $p\text{s} < 0.059$, Cliff's $\delta\text{s} > 0.20$), primarily because participants tended to make more omniscience errors for accessible MP_{E} problems that used *believe* as the epistemic verb: they generated errors on 63% of trials for such problems. The three-way interaction, however, was not significant.

Experiment 1b. Participants made omniscience errors on 50% of trials in Experiment 1b; 39 out of 43 participants produced at least one such error (binomial test, $p < .0001$, chance probability of 5%). They yielded such errors 54% of the time for accessible problems versus 46% of the time for inaccessible problems (Wilcoxon test, $z = 2.07$, $p = .038$, Cliff's $\delta = 0.14$). They made omniscience errors more than chance for both accessible and inaccessible problems (Wilcoxon tests, $z\text{s} > 5.42$, $p\text{s} < 0.001$, Cliff's $\delta\text{s} > 0.76$). There was no reliable difference between the percentages of errors they produced for the non-factive verb *thinks* compared to the factive *understands* (52% vs. 48%, Wilcoxon test, $z = 1.37$, $p = .172$, Cliff's $\delta = 0.07$), but there was a difference between MP_{E} and MT_{E} problems: the former yielded more errors (56% vs. 43%; Wilcoxon test, $z = 2.51$, $p = .012$, Cliff's $\delta = 0.20$).

There was a marginal interaction between participants' production

Table 2
Percentages of different sorts of error across Experiments 1a-e. Italics provide an example problem as well as examples of responses that correspond to that problem.

	Omniscience errors	Errors of omission	Logical errors	Nonsensical responses	Accurate responses
	Ex., Ari knows that if a jar is in the walkway, then the password is 'pear'. A jar is in the walkway.				
Experiment	*Ari knows that the password is 'pear.'*	*Nothing follows.*	*The password is not 'pear.'*	*The knows password Ari.*	*The password is 'pear.'*
1a	44%	5%	12%	8%	32%
1b	50%	4%	7%	11%	28%
1c	39%	3%	14%	11%	33%
1d	27%	3%	12%	16%	44%
1e	17%	10%	22%	3%	57%

of omniscience errors based on the accessibility of information and the problem's structure: accessible MP_E problems yielded errors on 64% of trials compared to much fewer errors (< 49%) for the other categories (Wilcoxon test, $z = 1.89$, $p = .059$, Cliff's $\delta = 0.21$); no other two-way interaction was significant. However, the three-way interaction was reliable (Wilcoxon test, $z = 2.23$, $p = .026$, Cliff's $\delta = 0.21$).

Experiment 1c. As in the previous studies, participants made omniscience errors more often than any other response in Experiment 1c, i.e., on 39% of the trials: 37 out of 43 participants made at least one such error (binomial test, $p < .0001$, chance probability of 5%). They yielded omniscience errors 44% of the time for accessible problems versus 34% of the time for inaccessible problems (Wilcoxon test, $z = 1.88$, $p = .060$, Cliff's $\delta = 0.18$); did so more often for the non-factive *believe* than the factive *know* (45% vs. 33%, Wilcoxon test, $z = 3.12$, $p = .001$, Cliff's $\delta = 0.24$); and did so more often for MP_E than MT_E structures (45% vs. 33%, Wilcoxon test, $z = 2.93$, $p = .003$, Cliff's $\delta = 0.20$). Participants produced omniscience errors significantly more often than chance (Wilcoxon tests, $zs > 5.24$, $ps < 0.001$, Cliff's $\delta s > 0.69$).

No two-way interactions were reliable in Experiment 1c, however a significant three-way interaction revealed the same pattern that previous studies showed: participants made more errors for accessible MP_E problems that used *believe* as the epistemic verb (Wilcoxon test, $z = 2.21$, $p = .027$, Cliff's $\delta = 0.26$).

Experiment 1d. Participants generated omniscience errors on 27% of the problems in Experiment 1d; 26 out of 45 participants made such errors on at least one trial (binomial test, $p < .001$, chance probability of 5%). The result is particularly striking, because participants were instructed on each problem that the agent doesn't know the information relevant to making any sort of inference. This experiment included only inaccessible problems. Participants' tendency to generate errors didn't differ reliably based on the epistemic verb in the problem (*know*: 26% errors vs. *believe*: 28%, Wilcoxon test, $z = 1.04$, $p = .30$, Cliff's $\delta = 0.02$). And there was no difference as a function of problem structure (MP_E: 26% vs. MT_E: 28%; Wilcoxon test, $z = 0.82$, $p = .41$, Cliff's $\delta = 0.03$). Likewise, the interaction between the two was not reliable (Wilcoxon test, $z = 0.13$, $p = .90$, Cliff's $\delta = 0.04$). In Experiment 1d, unlike the previous studies, the most common response was to produce a correct conclusion, which depended on the problem itself – e.g., "nothing follows" was the correct conclusion to (30) while "the password is pear" was the correct conclusion to (31). Participants produced more correct conclusions than omniscience errors (44% vs. 27%; Wilcoxon test, $z = 5.54$, $p < .001$, Cliff's $\delta = 0.17$). Nevertheless, they produced omniscience errors significantly more than chance (Wilcoxon test, $z = 4.97$, $p < .001$, Cliff's $\delta = 0.58$).

Experiment 1e. Participants generated omniscience errors on 17% of the problems in Experiment 1e; 30 out of 55 participants made such errors on at least one trial (binomial test, $p < .001$, chance probability of 5%). Similar to Experiment 1d, participants were instructed on each problem that the agent doesn't know the information relevant to making any sort of inference; the study included inaccessible problems only. Participants' tendency to generate errors didn't differ reliably based on the epistemic verb in the problem (*know*: 18% errors vs. *believe*: 17%, Wilcoxon test, $z = 0.32$, $p = .75$, Cliff's $\delta = 0.01$). And there was no difference as a function of problem structure (MP_E: 17% vs. MT_E: 18%; Wilcoxon test, $z = 0.21$, $p = .83$, Cliff's $\delta = 0.05$). They produced more omniscience errors when asked to consider what the agent in the problem would conclude rather than themselves (*agent:* 21% vs. *participant:* 13%, Wilcoxon test, $z = 2.69$, $p = .007$, Cliff's $\delta = 0.19$). No interactions were significant.

In Experiment 1e, like Experiment 1d, the most common response was to produce a correct conclusion. Participants produced more correct conclusions than omniscience errors (57% vs. 17%; Wilcoxon test, $z = 4.30$, $p < .001$, Cliff's $\delta = 0.67$). Nevertheless, they produced errors significantly more than chance (Wilcoxon test, $z = 5.36$, $p < .001$, Cliff's $\delta = 0.54$).

3.1.1.3. Results. Participants produced omniscience errors in Experiments 1a-e: such errors were the most common response they constructed for every study except Experiments 1d and 1e. Participants tended to generate more such errors for *believe* than for *know* (Experiments 1a and 1c); descriptions of inaccessible evidence appeared to reduce the errors (Experiments 1a-c). They generated errors for both of the two structures – MP_E and MT_E – used in the studies. And across every condition of every study, participants made errors significantly above chance. The studies were designed to rule out alternative explanations for the effect: Experiment 1b ruled out the possibility that the effect comes from generating responses that matched the verbs in the premises; Experiment 1c ruled out the possibility that the effect came from a pragmatic framing that encouraged reasoners to attribute the facts to the agent described in the premises. Experiments 1d and 1e should have, in theory, eliminated all omniscience errors, because it presented participants with explicit information about what the agent doesn't know. Participants made errors, nevertheless.

One limitation of Experiments 1a-e is its reliance on the sentence completion task we designed (see Appendix B). This task sought to give participants flexibility in producing conclusions in an unbiased manner while ensuring that their responses were relevant to the task of considering the mental states of the agents in each problem. And it yielded results that were systematic and consistent across the different studies (see Appendix C for additional analyses of response systematicity). Yet it is still possible that the task and its interface introduced subtle biases and forced participants to consider conclusions that they otherwise might not. One obvious limitation of the task is that, by presenting participants with sentence fragments that concerned mental state verbs, the experiments may have encouraged them to use those fragments. A more subtle limitation of the task and its interface is that it permitted participants to consider conclusions piecemeal; participant could press a "reset" button to clear any sentence they had constructed. This functionality allowed them to correct any errors or mistaken clicks they might have made, but participants may have used it to deliberate about the conclusions they provided. To address these limitations, Experiment 2 used a design and materials similar to the previous studies, but it asked participants to type out their responses to problems such as (30−32).

3.1.2. Experiment 2

Experiment 2 was almost identical to Experiment 1a – it adopted the same design and materials as the previous study and presented participants with problems akin to (30–32). But instead of asking participants to respond by constructing sentences using the interface introduced earlier, participants had to type out responses to the question, "What, if anything, follows?" Participants could respond however they wished, and as we will show, many of them chose to respond in a way that did not reference any mental states whatsoever. Indeed, their natural responses were heterogeneous enough that we could not analyze them using the same coding rubric devised for the earlier studies. We therefore developed a separate coding rubric, which we describe below, and used it to isolate and analyze any omniscience errors that participants produced.

3.1.2.1. Method. Participants. 69 participants in Experiment 2 volunteered through Amazon Mechanical Turk for monetary compensation (mean age = 39.26 years; 35 females, 33 males, 1 prefer not to answer); 14 failed attention checks and were excluded from analysis, which yielded data from 55 participants.

Design, materials, and procedure. Experiment 2 presented participants 16 separate problems and 2 attention checks using the same materials and problem structures as in Experiment 1a. It therefore manipulated the epistemic verb in each problem, the problem structure, and the accessibility of information, and accordingly adopted a 2 (*know* vs. *believe*) x 2 (MP_E vs. MT_E) x 2 (accessible vs. inaccessible) repeated measures design. The names, locations, and passwords in the problems were all randomized, as was the order in which participants carried out the problems. After presenting a problem such as (32) above, participants typed their answers into a box provided on the screen. They were required to type out an answer at least 7 characters long. The study incorporated two attention checks, which were designed differently than those used in the previous studies: attention check trials looked exactly like the other trials, except that they asked participants to type out a string of three repeated words (i.e., “strawberry strawberry strawberry”). Participants could take as long as they needed to type out responses.

Coding rubric and interrater reliability. The first and second authors developed a coding rubric to annotate participants' typed responses. Since Experiment 2 permitted participants to respond in any way they wished, responses contained sentence fragments, ungrammatical sentences, misspellings, and other forms of aberrant output. The rubric therefore sought to capture as much information from those responses as possible. It coded participants' responses along seven categories:

a. *Agentic* responses referred directly to an agent by name or pronoun (e.g., “Alex knows that the password is ‘kiwi’”).
b. *Epistemic* responses referenced a knowledge or belief state, or else a lack of knowledge or belief (e.g., “Krish will find out whether or not the password is ‘watermelon’”).
c. *Hedged* responses mention the possibility of an epistemic state, or a state of things in the world, without committing to them. Hedged responses could be disjunctive in nature by mentioning both a state and its negation as equal possibilities (e.g., “A broom is not in the office or the logic is wrong” and “A padlock is likely not in the museum”)
d. *Invalid* deductions described logically incorrect judgments and depended on the type of problem structure used (e.g., “nothing follows” is an invalid deduction for an MP_E problem such as (31) above)
e. *Valid* deductions were deductively correct conclusions and depended on the type of problem
f. *Nonsense* responses were unintelligible, incoherent, or else a string of text unrelated to the problem
g. *Omniscience errors* were erroneous conclusions that concern an agent possessing knowledge or beliefs about the state of the world, or else acting on the world in a way that reflects that mental state (e.g. “Jamile inputs the password ‘cherry’”)

Each response could reflect various combination of the seven codes, e.g., responses could have been both agentic and epistemic, and because participants could have provided multiple responses (though this was rare). Alternatively, responses could reflect none of the codes above – such responses indicated situations in which, e.g., a participant elaborated on the premises (e.g., “Demetrice is still looking in the library”) or else those that instructed the agent in some fashion (e.g., “change the password”). The first and second authors coded 10% of the responses together on all 7 categories, assessed discrepancies and interrater reliability, refined the rubric, and then reassessed reliability. Table 3 provides interrater reliability results for the initial and final versions of the rubric. The first author used the refined rubric to code the remainder of the data.

Table 3
Interrater reliability results on data coded jointly by the first and second authors along seven categories relevant to the analysis of omniscience errors; results are shown for both the initial version of the rubric and its final version.

Coding category	% of responses	Cohen's κs	
		Initial version	Final version
Agentic responses	35%	0.78	0.91
Epistemic responses	18%	0.56	0.95
Hedged responses	4%	0.94	1.0
Invalid deductions	38%	0.65	1.0
Valid deductions	45%	0.57	0.85
Nonsense responses	< 1%	0.65	0.66
Omniscience errors	13%	0.67	1.0

3.1.2.2. Results and discussion. For brevity, we assess here only participants' tendencies to produce omniscience errors; full analyses across all coding categories as a function of all the manipulations in the study are available online. Participants produced omniscience errors on 13% of trials in Experiment 2, and neither the verb in the premises nor the accessibility of information affected their tendency to do so (Wilcoxon tests, zs < 0.72, ps > 0.46, Cliff's δs < 0.07). Likewise, none of the interactions affected the pattern of their omniscience errors (Wilcoxon tests, zs < 1.16, ps > 0.24, Cliff's δs < 0.09). However, one factor vastly affected error production: the logical structure of the problem. Participants generated omniscience errors 22% of the time for MP_E problems but only 5% of the time for MT_E problems (Wilcoxon test, $z = 4.25$, $p < .001$, Cliff's $\delta = 0.31$). The principle of mental state tags alone cannot account for this difference; it occurred in some, but not all, of the previous studies. One reason it may have occurred in Experiment 2 is because, as ancillary analyses show, participants produced more valid deductions for MT_E vs. MP_E problems (39% vs. 32%) and fewer invalid deductions for MT_E vs. MP_E problems (41% vs. 49%). In non-epistemic contexts, such as (2) above, participants tend to infer that “nothing follows” from modus tollens problems. In the present experiment, such a response is an error when MT_E problems concern factive verbs – but it is sensible for non-factive verbs. Participants may have produced more “nothing follows” responses for all types of problems in the study, and doing so may have reduced the opportunity for them to make omniscience errors.

As the study shows, participants produced omniscience errors systematically. They did so less often than when they used the sentence completion task introduced in the previous studies. Nevertheless, the results validate the principle of mental state tags, which expects that participants should make omniscience errors as a consequence of the misapplication of those tags.

3.1.3. Experiment 3

Experiment 3 tested participants' tendencies to engage in meta-epistemic reasoning, that is, how they reason about one agent's mental states about another agent's mental states. Consider this problem:

33. Amari and Riley come across a particular animal [at a wildlife preserve].
    Amari believes that Riley knows that the animal is a buzzard.
    Does it follow that the animal is a buzzard?

In (33), Amari's belief concerns, not the relevant facts, but rather Riley's knowledge. The tag principle holds that people should generate the following model:

```
animal is buzzard   [Riley]     (Amari)
```

which includes a factive tag associating Riley with the mental state and a non-factive tag associating Amari with the model representing Riley's mental state. The [Riley] tag is factive, and it forces a presupposition that the animal is a buzzard – and so reasoners should be highly likely to say "yes" to (33). The question in (33) could be replaced by others, e.g.,

34a. Does it follow that Riley knows the animal is a buzzard?
b. Does it follow that Riley believes the animal is a buzzard?
c. Does it follow that Amari knows the animal is a buzzard?
d. Does it follow that Amari believes the animal is a buzzard?

Let us analyze what the theory predicts for each such question: because (33) should cause reasoners to represent Riley's mental state as a state of knowledge, a scan of the model above should cause them to infer that Riley knows that the animal is a buzzard. But such a conclusion is an error: Amari merely believes that Riley knows what the animal is, but Amari could be mistaken. Yet, if Riley knows the animal is a buzzard, she also believes it: a factive tag permits non-factive conclusions. Hence, reasoners should say "yes" to (34a) and (34b). The theory likewise predicts that they should say "yes" to other non-factive conclusions, such as:

35a. Riley thinks the animal is a buzzard.
b. Riley feels as though the animal is a buzzard.
c. Riley claims that the animal is a buzzard.

and so on.

The tag representing Amari's mental state is non-factive, and so it cannot be used to infer anything about what Amari knows: reasoners should reject (34c). But, because the model represents a tag associating Amari's mental state to the proposition that the animal is a buzzard, reasoners may not reject (34c) fully. In contrast, they should reject these conclusions fully:

36a. Amari knows the animal is a cockatiel.
b. Amari believes the animal is a duck.

and so on, because the model above does not directly represent the information in (36a). More succinctly: reasoners should say "yes" to (34c) on some non-zero minority of trials, whereas they should rarely if ever accept (36a). Finally, consider (34d): the model contains a non-factive tag associating Amari's belief with the animal being a buzzard. Hence, they should say "yes" to (34d) more often than not.

The tag principle makes analogous predictions for problems that incorporate different versions of the second premise in (33), e.g.,

37a. Amari knows that Riley knows that the animal is a buzzard.
b. Amari knows that Riley believes that the animal is a buzzard.
c. Amari believes that Riley believes that the animal is a buzzard.

For brevity we omit an analysis for each such problem, but we highlight two central patterns. (37a) should cause reasoners to say "yes" most of the time to each of the questions in (34a-d). In contrast, (37c) should cause reasoners to reject, e.g., that Riley knows that the animal is a buzzard (34a). But it should cause them to accept that Amari believes that the animal is a buzzard – which constitutes a flaw in reasoning, because Amari can maintain beliefs about Riley's beliefs without believing them himself.

Experiment 3 used problems such as (33) and variations such as (37a-c) to test the detailed predictions of the tag principle.

3.1.3.1. Method. Participants. 52 participants in volunteered through Amazon Mechanical Turk for monetary compensation (mean age = 36.96 years; 29 females, 23 males); 16 failed attention checks and were excluded from analysis, which yielded data from 36 participants.

Design, materials, and procedure. Experiment 3 instructed participants to imagine pairs of visitors at a wildlife park as they think about animals they may encounter. Problems were similar to (33) above: the first premise introduced two agents, and the second described one agent's mental state about another agent's mental state using one of four separate structures:

X knows that Y knows that P.
X knows that Y believes that P.
X believes that Y knows that P.
X believes that Y believes that P.

Participants then assessed whether a particular conclusion followed of necessity by clicking buttons marked "Yes" or "No". The conclusions were in one of five separate structures:

P.
X knows that P.
X believes that P.
Y knows that P.
Y believes that P.

Participants saw every combination of these two sorts of structures once and so Experiment 3 reflected a 4 (premise structure, e.g., *X knows that Y knows that P*) x 5 (conclusion structure, e.g., *P*) repeated measures design. They carried out each such combination once and the study used two separate attention checks, and so each participant carried out 22 problems. An attention check looked entirely similar to regular problems, but instead of asking participants to decide whether the conclusion followed, it instructed participants: "For this trial simply click this plus button [+] to continue" and provided a button on the screen that was formatted analogously to the "Yes" and "No" buttons. Participants passed the check by clicking the "+" button instead of the others. The experiment randomized the positions of the "Yes" and "No" buttons.

3.1.3.2. Results and discussion. Table 4 shows the percentages of "yes" responses as a function of the premise and conclusion structures. We subjected the data to nonparametric analyses of variance, and we

Table 4
Percentages of "yes" responses on Experiment 3 as a function of the structures of the premises and conclusions in Experiment 3. Bolded cells denote when participants accepted the conclusions significantly more than chance (50%).

Premise	Conclusion				
	P	*Y knows that P*	*Y believes that P*	*X knows that P*	*X believes that P*
X knows that Y knows that P	**89**	**92**	**69**	**67**	**67**
X knows that Y believes that P	53	47	**75**	44	**69**
X believes that Y knows that P	**81**	**83**	**69**	42	**69**
X believes that Y believes that P	56	53	**78**	39	**69**

highlight main effects and the interaction in brief: responses differed reliably as a function of the premise structure (Friedman test, $\chi^2 = 13.21$, $p < .005$) and conclusion structure (Friedman test $\chi^2 = 19.75$, $p < .001$), and these two factors marginally interacted with each other (Friedman test, $\chi^2 = 7.85$, $p = .10$).

To test whether these patterns accord with the predictions of the tag principle, we subjected the data from each cell in Table 4 to pairwise nonparametric analyses against chance performance (50%), and *p*-values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure. The table highlights those results. It shows that for *X knows that Y knows that P*, participants reliably accepted all five conclusions more than chance (Wilcoxon tests, $zs > 1.99$, $ps < 0.045$, Cliff's $\delta s > 0.32$). For *X knows that Y believes that P* participants accepted only the conclusions that *Y believes that P* and that *X believes that P* (Wilcoxon tests, $zs > 2.32$, $ps < 0.03$, Cliff's $\delta s > 0.37$). The former conclusion is sensible; the latter conclusion is an error that the theory predicts. For *X believes that Y knows that P*, participants accepted all the conclusions (Wilcoxon tests, $zs > 2.32$, $ps < 0.03$, Cliff's $\delta s > 0.37$) except that *X knows that P*. And for *X believes that Y believes that P*, participants accepted the conclusions that *Y believes that P* and that *X believes that P* (Wilcoxon tests, $zs > 2.33$, $ps < 0.03$, Cliff's $\delta s > 0.37$), both of which are also predicted errors. These patterns are distinct from the omniscience errors described in Experiments 1 and 2; the tag principle accounts for each of these errors in meta-epistemic reasoning.

3.1.4. *Interim summary*

Experiments 1a-e and Experiment 2 corroborated the tag principle's prediction that reasoners should make systematic mental state reasoning errors: if they can infer a deductive conclusion for premises that describe an agent's mental states, they infer that the agent knows or believes that conclusion and thereby ascribe undue omniscience to the agent. Experiment 3 showed that they engage in meta-epistemic reasoning, i.e., reasoning about second-order knowledge and belief. This study, too, revealed participants' systematic errors. Tags therefore allow reasoners to categorize and track mental states – but they allow reasoners to draw conclusions that do not follow of necessity.

3.2. *Testing the principle of consolidation*

How do you make inferences about what follows from an agent's knowledge or belief? It may be impossible to anticipate another person's actions without inferring what they will reason from the information they acquire. Your conclusions about what they might conclude may be quite distinct from what you yourself conclude about a situation. To explain these patterns, the principle of consolidation assumes that people reason about others' inferences by combining models that bear similar mental state tags. Consider what you might conclude from (6) above, which we repeat here:

6. Olga knows that if there's an ace in the deck, then there's a queen in the deck.
   Olga knows that there's an ace in the deck.

The principle of consolidation proposes that you build models of these two mental states and combine them, and that the process of doing so preserves their tags. Hence, it predicts that the most common conclusion people should draw is that *Olga knows there's a queen in the deck* (7a), even though – because the two statements above presuppose the truth of their complements – it's just as reasonable to conclude that there's a queen in the deck (7b). If the second premise in (6) used the verb *believe* instead of *know*, the principle predicts that people should attribute a mental state of belief to Olga. Experiment 4 tested both of these predictions.

3.2.1. *Experiment 4*

Experiment 4 presented problems such as the following:

38. Sammy notices something moving near some trees.
    Sammy knows that if it's July, then it's a fox.
    Sammy knows that it's July.

Participants had to respond through a variant of the sentence completion task used in Experiments 1a-e (see Appendix B), which presented them with a list of phrases that appeared within boxes on the screen similar to the interface shown in Fig. A1. The list of phrases corresponded to the following options relevant to (38) above: *Sammy / knows that / believes that / it's a fox / nothing follows*. Participants dragged one or more of the boxes to an area designed on the right of the screen where they could progressively populate a sentence, such as: *Sammy believes that it's a fox.*

3.2.1.1. Method. Participants. 49 participants in Experiment 4 volunteered through Amazon Mechanical Turk for monetary compensation. (mean age = 40.08 years; 26 females, 23 males); 4 failed attention checks and were excluded from analysis, which yielded data from 45 participants.

Design and materials. Participants completed 16 problems and 2 attention checks. Problems consisted of three premises: the first premise established that an individual "notices something moving near some trees" in a particular wildlife preserve. The second and third premises presented a conditional (*if P then Q*) and a categorical statement (*P*) couched in terms of an individual's mental state, e.g., (*Sammy knows that if P then Q*), where *P* concerned some named interval of time (e.g., *it's July* or *it's Monday*) and *Q* concerned the presence of some animal (e.g., *it's a fox*). The experiment manipulated the verbs in the first and second premises, which could be either *know* or *believe* – hence, the study reflected a 2 × 2 repeated measures design. The contents of the problem were drawn from a list of 20 names, 20 intervals, and 20 animals in a manner that ensured no repetitions across the study. The experiment randomized the order in which the second and third premises appeared on the screen, and it randomized the order of the words in the word list. It also randomized the order of the individual problems, which ensured that no participant received the same combination of problems and materials in the same order. As in Experiments 1a-e, attention check trials asked participants to create an ungrammatical sentence (e.g., "believes that knows that it's a fox").

Procedure. The experiment instructed participants on how to use the drag-and-drop variant of the interface described in Appendix B.

Table 5
Percentages of responses along the five coded categories in Experiment 4 as a function of the four problems participants saw in the study. Bolded cells reflect the most frequent response.

Conditional premise	Categorical premise	Response				
		Q	*X knows Q*	*X believes Q*	*Nothing follows*	*Nonsense*
X knows that *if P then Q*	X knows that *P*	10	**77**	7	5	1
X knows that *if P then Q*	X believes that *P*	3	12	**82**	3	0
X believes that *if P then Q*	X knows that *P*	4	15	**76**	3	1
X believes that *if P then Q*	X believes that *P*	3	3	**90**	3	0

Participants learned how to populate a sentence using materials and verbs that never appeared in the experiment proper; they learned explicitly that they could produce any grammatical response they chose, including options such as “it’s a fox” and “nothing follows” (each of which required them to drag only one box to the assigned area) as well as longer sentences such as “Sammy knows that nothing follows” and “Sammy believes that it’s a fox”. After practicing forming two sentences in the interface, participants carried out the remainder of the problems in the study. Each problem instructed them to “drag words to fill in the blank to indicate what follows” from problems such as (38) above.

Coding. We categorized participants’ responses on whether they reflected the following three categories: *Q*, *X knows Q*, *X believes Q*. If they fell into none of these categories, we assessed whether they responded that nothing follows, or else whether they were nonsensical.

Open science. Experimental code, materials, data, coding rubric, and analyses are available at https://osf.io/s24ak/.

3.2.1.2. Results and discussion. Table 5 shows the percentages of the responses for the four types of problems in Experiment 4 along the five coded categories. Given the multinomial nature of the responses, we examine relevant pairwise comparisons for brevity. Across the study, people tended to draw conclusions about agents’ mental states more often than conclusions about presuppositions ($Q = 5\%$ vs. *X knows* $Q = 27\%$; Wilcoxon test, $z = 4.58$, $p < .001$, Cliff’s $\delta = 0.35$; $Q = 5\%$ vs. *X believes* $Q = 64\%$; Wilcoxon test, $z = 5.79$, $p < .001$, Cliff’s $\delta = 0.37$), which corroborates both the principle of consolidation and suggests that reasoners combine existing models in a manner that retains their epistemic tags. Likewise, participants responded that *X believes Q* far more often than they responded *X knows Q* (64% vs. 27%, Wilcoxon test, $z = 5.75$, $p < .001$, Cliff’s $\delta = 0.67$), which the consolidation principle predicts. They produced *X knows Q* as their most frequent response only when both of the epistemic verbs in the problem were both *know*; they preferred to produce responses such as *X knows Q* over *Q* for such problems (77% vs. 10%, Wilcoxon test, $z = 4.59$, $p < .001$, Cliff’s $\delta = 0.73$), even though both conclusions are sensible. Both of these patterns validate the predictions of the principle of consolidation.

For the three other types of problems, participants tended to infer that *X believes Q*; they produced such responses more often than *X knows Q* (83% vs. 10%, Wilcoxon test, $z = 5.88$, $p < .001$, Cliff’s $\delta = 0.79$), which, too, corroborates the principle of consolidation. They produced responses such as *X knows Q* more often when at least one verb in the problem was *know* compared to problems such as *X believes that if P then Q / X believes that P* (35% vs. 3%, Wilcoxon test, $z = 5.81$, $p < .001$, Cliff’s $\delta = 0.21$).

The results corroborate the principle of consolidation and suggest that reasoners combine models and retain their epistemic tags, which can bias the conclusions that they infer.

3.3. Testing the principle of alternatives

The principle of alternatives explains how people integrate online knowledge, i.e., presuppositions that come from expressions of language used to describe what agents know, and offline knowledge, i.e., relational and conceptual facts stored in long-term memory. The model theory specifies algorithms for combining two separate models (Johnson-Laird & Khemlani, 2023), and these same algorithms apply towards integrating models of online and offline knowledge and reducing the number of alternative possibilities reasoners can consider. For non-factive verbs, such as *think*, *assume*, *guess*, and *believe,* reasoners construct a model of an individual’s beliefs (see Harner & Khemlani, 2022) without presupposing any fact, and hence online beliefs, such as in (29) above, permit reasoners to consider many different alternatives.

This account of how people process expressions of knowledge predicts a novel pattern of reasoning: the effects of modulation should not occur in scenarios expressing beliefs instead of knowledge. Consider this inference:

39. Loma knows that if it’s an animal, then it’s a wolf.
    It’s not an animal.
    Does it follow that it’s not a wolf?

Reasoners should infer that it’s not a wolf and therefore make an epistemic analog of a denial of the antecedent inference (hence, we abbreviate this problem structure as DA_E), because the conditional in (39) expresses a state of knowledge held by Loma. The presupposition is the conditional itself, which is not a fact of the world but rather a set of possibilities. Offline knowledge modulates the conditional by ruling out two alternatives: one in which it’s an animal but not a wolf, and another in which it’s a wolf but not an animal. And so people should be less prone to drawing a DA_E inference for unmodulated conditionals (e.g., *Loma knows that if it’s an animal, then it’s hidden*) because these conditionals don’t rule out those same alternatives. But the effect of modulation should disappear for non-factive expressions of belief, as in:

39′. Loma believes that if it’s an animal, then it’s a wolf.
It’s not an animal.
Does it follow that it’s not a wolf?

The epistemic verb *believe* makes no presupposition, so it permits reasoners to consider many different alternatives. The following experiment tested and confirmed this interaction.

3.3.1. Experiment 5

Experiment 5 tested the interaction between online and offline knowledge: for expressions that concern the factive epistemic verb *know,* reasoners should exhibit a modulation effect, i.e., they should be more likely to make inferences akin to an affirmation of the consequent (AC) and a denial of the antecedent (DA) for modulated than unmodulated conditionals. For expressions that concern the epistemic verb *believe,* the theory predicts no difference between modulated and unmodulated conditionals.

Participants in the study saw problems such as this one (which we describe as an epistemic analog of an AC problem, and so abbreviate it AC_E):

40. Devon knows that if it’s cloudy, then it’s a warthog.
    Devon knows that it’s a warthog.
    Is it cloudy?

The problem matches the following logical structure:

41. If P, then Q.
    Q.
    Does it follow that P?

though it embeds the premises in statements that ascribe knowledge to a particular individual. The principle of alternatives predicts that people should reject (4) for the unmodulated conditional above, but that they should accept it for modulated conditionals such as, “if it’s an animal, then it’s a warthog.” And it predicts that this effect of modulation should hold for factive verbs (e.g., *knows*) but not for non-factive verbs (e.g., *believes*).

3.3.1.1. Method. Participants. Experiment 5 recruited 60 healthy members of the general North American public through the Cloud Research online platform (29 females, 31 males, 0 other/prefer not to say; mean age = 36.77, age range = 22–62) and compensated them $1.50 for a study that lasted less than 6 min; 6 participants were excluded from statistical analysis for failing to meet attention check criteria. 40 out of the remaining 54 participants had received no prior instruction in symbolic logic.

Design and materials. Participants completed 12 problems in total. Each problem consisted of three premises: the first premise introduced an agent and an observation (e.g., "[Agent] notices something in the distance"); a second premise stipulated that the agent possessed conditional knowledge linking a state of affairs to an animal (e.g., "[Agent] knows that [if P then Q]"); and a third premise stipulated that the agent possessed categorical knowledge of one of the clauses of the conditional (e.g., "[Agent] knows that [Q]."). Participants then assessed whether a particular conclusion followed from the given premises (e.g., equivalent to, "Is it the case that [P]?") by registering their response on buttons marked "Yes", "No", and "I'm not sure".

Experiment 5's primary manipulation concerned semantic modulation, that is, whether the conditional premise described an *if*-clause and a *then*-clause that prohibited certain possibilities. The experiment constructed modulated conditionals by using the following *if*-clause: "it is an animal". For example: "…if it is an animal, then it's an ostrich." Unmodulated conditionals concerned a weather condition, e.g., "…if it's cloudy, then it's an ostrich." The materials in each unmodulated problem came randomly drawn from a pool of weather conditions (e.g., "it's cloudy") and the materials in each modulated problem were drawn from a pool of animals (e.g., "ostrich"). As the model theory predicts, the difference between the two is that reasoners' knowledge of various animals suppresses the consideration of any possibility in which it's an ostrich but not an animal (e.g., *not-P and Q*), whereas no such suppression occurs for scenarios in which it's an ostrich but not cloudy.

A secondary manipulation concerned the epistemic verb used to stipulate online knowledge. That is, half the problems concerned an agent's knowledge (e.g., "Devon knows that if…") and the other half concerned an agent's belief (e.g., "Devon believes that if…"). To vary both the presentation of the materials as well as the structures of the problems, the study also manipulated whether the problem structure reflected a mental state analog of an affirmation of the consequent inference (AC_E) or a denial of the antecedent inference (DA_E). In their classical form, AC and DA inferences are logically invalid, but compelling (see, e.g., Barrouillet, Gauffroy, & Lecas, 2008; Evans, 1993; Oberauer, 2006; Singmann et al., 2014). The model theory predicts that epistemic verbs should affect the endorsement of AC_E and DA_E inferences. The experiment randomized the names used for the agents, the materials assigned to the conditions, the order of the problems, and the positions of the response buttons. Two attention check trials were similar in all respects to the 8 other problems in the experiment except that a separate button appeared on the screen for participants to press to indicate that they were paying attention. We excluded participants who missed both attention check trials from subsequent analyses. In addition, two "interpretation" trials were included to verify understanding of *believe* and *know*. These trials consisted of an agent in a location and a weather event taking place in a different location. These were included to get an understanding of how participants had interpreted each epistemic verb.

Open science. The experimental code, materials, data, and statistical analyses are available through the Open Science Framework (https://osf.io/36b9u/), as are preregistrations for all analyses.

3.3.1.2. Results and discussion. Participants in the study endorsed inferences (e.g., accepted AC_E or DA_E inferences) reliably more often for modulated problems than for unmodulated problems (60% vs. 36%, Wilcoxon test, $z = 3.29$, $p < .001$, Cliff's $\delta = 0.48$), a pattern that corroborates the principle of alternatives, and one that mirrors modulation affects observed in other domains (Quelhas et al., 2019). Participants endorsed AC_E and DA_E inferences more often when the epistemic verb was factive rather than non-factive (62% vs. 35%, Wilcoxon test, $z = 4.33$, $p < .001$, Cliff's $\delta = 0.43$). The structure of the problem, i.e., AC_E or DA_E, didn't affect their tendency to endorse inferences (49% vs. 47%, Wilcoxon test, $z = 0.52$, $p = .60$, Cliff's $\delta = 0.05$).

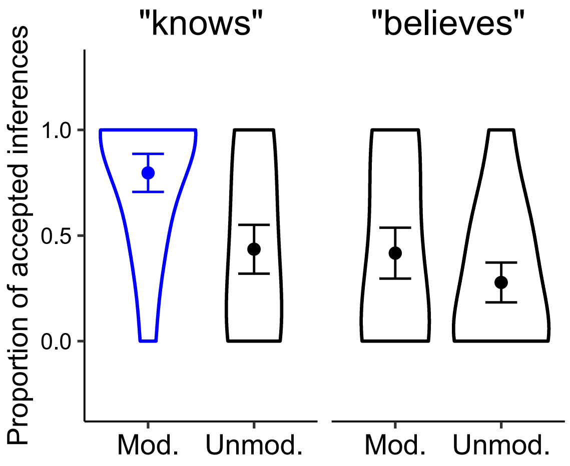

Fig. 2. Participants in Experiment 5 carried out problems of the form: *X [knows / believes] that if P then Q; Q is true; Does it follow that P?* The figure shows violin plots of the jittered proportions of accepted AC' or DA' inferences in the experiment as a function of whether the conditional (*if P then Q*) was modulated, and as a function of whether the epistemic verb was factive ("knows") or not ("believes") in each problem. Participants accepted inferences significantly more than chance only for modulated problems whose epistemic verb was factive (shown in **blue**) and not in any of the other conditions (shown in **black**). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The results yielded a reliable two-way interaction between modulation and factivity as predicted by the principle of modulation (see Fig. 2; Wilcoxon test, $z = 2.66$, $p = .007$, Cliff's $\delta = 0.27$): participants accepted modulated factives 80% of the time, and they accepted all other problems less than 44% of the time. The results likewise yielded a two-way interaction between the type of verb and the type of problem, i.e., AC_E vs. DA_E (Wilcoxon test, $z = 2.42$, $p = .02$, Cliff's $\delta = 0.19$). It yielded no other significant interactions. Data were subjected to a generalized mixed-model regression (GLMM) to control for participant- and material-wise random effects; the GLMM corroborated nonparametric analyses, i.e., it yielded a main effect of modulation ($B = 0.82$, $SE = 0.36$, $p = .02$), a main effect of factivity ($B = 2.37$, $SE = 0.42$, $p < .001$), and an interaction between modulation and factivity ($B = 1.4$, $SE = 0.55$, $p = .009$).

In sum, Experiment 5 revealed effects of modulation (an effect of offline background knowledge), of factivity (an effect of online knowledge ascription), and of the interaction between the two. The effects show that people do not reason based on the logical structure of premises, but rather on their meanings as embodied in models of possibilities. The effect is reinforced by online knowledge, but it can be reversed by online beliefs, since beliefs permit reasoners to consider alternative possibilities that they'd otherwise disregard.

4. General discussion

We present experiments that revealed four systematic epistemic reasoning patterns, which we summarize:

- Omniscience errors (Experiments 1a-e, 2): A reasoner concludes that *an agent knows Q is true* after learning that *P is true* and that the agent knows that *if P is true then Q is true*. The agent's knowledge of the conditional doesn't guarantee any knowledge about the truth or falsity of *C*.

- Meta-epistemic errors (Experiment 3): A reasoner concludes that an agent X believes P after learning that *X believes that another agent Y believes that P*. Belief about another agent's mental states does not permit any conclusion about what X believes.
- Consolidation errors (Experiment 4): A reasoner concludes that an agent X, who believes if P then Q and knows that P, also as a consequence knows Q. This is faulty because an agent can be false about their conditional belief.
- Modulation interactions (Experiment 5): A reasoner integrates semantic knowledge to modulate inferences for factive verbs but not non-factive verbs.

If the errors above were rampant, then humans would have difficulty engaging in anything more complex than rudimentary epistemic reasoning. Mental state reasoning errors may, in fact, be very rare in real life – communication would be impossible if they were very common. Perhaps theories of epistemic reasoning do not need to bother themselves with explaining such kinks and glitches in the reasoning process and should focus instead on capturing the many sensible inferences that reasoners make. Perhaps the errors above occur only in prescribed laboratory settings; perhaps they have no real correspondence to how people engage in epistemic reasoning in daily life.

Yet, as the studies show, omniscience errors occurred ~15% of the time on Experiment 2, and consolidation errors occurred ~20% of the time on Experiment 4. The errors reasoners made were not arbitrary; the tasks used in each of the studies were designed to promote reflection, and the answers people provided were ones they generated for themselves – so they likely did not occur by chance.

We developed a theory to explain why and how people make mental state inferences, both optimal and suboptimal. The theory addresses a deficit in the literature, because no alternative account explains epistemic reasoning at the algorithmic level of analysis: no theory proposes the structures of the mental representations that underlie mental state reasoning or the cognitive processes involved in tracking and maintaining those structures, so no existing theory can explain how such processes break down. The theory argues that people construct small-scale mental possibilities – mental models – to reason about mental states. Its central intuition is that *knowing* something prevents a reasoner from constructing and considering certain possible states of the world: to *know* that the current King of England is Charles III is to suppress any consideration of a situation in which somebody else is the King. Mental models mimic the structure of what they represent, and reasoners can hold only a limited number of models in working memory at any given moment. The theory posits three main principles of reasoning about mental states:

- reasoners tag possibilities with mental state information, and do so recursively when appropriate;
- they draw conclusions through a consolidation process, where they integrate the information contained in similarly tagged models;
- they search for alternative possibilities with respect to those tags (see Table 1 above).

The model theory does not predict widespread fallacious reasoning: in fact, it provides an account of how people competently reason about mental states, and it does so without appealing to cognitively implausible formal framework, such as epistemic or doxastic logic (see Johnson-Laird, Byrne, & Khemlani, 2024). It provides an explanation of the bookkeeping machinery reasoners use to keep their own mental states separate from others' (Experiments 3 and 4). It likewise shows how people can integrate background semantic knowledge into their representations of what agents know and believe (Experiment 5), and how those integration processes modulate the inferences they make. It explains how naïve individuals can deploy significant cognitive resources to make optimal mental state inferences: reasoners perform well when they consider the consequences that follow when one or more beliefs are mistakenBut the mental process of enumerating all of the possibilities consistent with both accurate and fallacious beliefs can run into processing and memory constraints, and so it explains reasoning errors as a consequence of the shortcuts reasoners take to avoid such protracted deliberation.

The experiments above test the most critical predictions of the theory – but the theory accounts for reasoning patterns that studies have yet to investigate. For instance, the theory's principle of consolidation posits that the verb *believe* permits the consideration of alternative possibilities, whereas *knows* prevents it – and so it explains many additional patterns about how people combine and coalesce explicit belief states. We illustrate one such pattern: when a non-factive enters into a set of premises, people can consider alternatives. In cases involving multiple premises, e.g.,

42. Ali knows that X or Y or Z is true.
    Ali believes X is false.
    Ali believes Y is false.

the principle predicts that people should conclude (sensibly): *Ali believes Z is true*. So, too, for this set of premises, which is the same as above but uses *know* instead of *believe* in the second premise:

42′. Ali knows that X or Y or Z is true.
Ali knows X is false.
Ali believes Y is false.

But, for this set of premises in which *know* is used throughout, participants should infer that *Ali knows that Z is true*, not just that she believes it:

42″. Ali knows that X or Y or Z is true.
Ali knows X is false.
Ali knows Y is false.

The pattern is not a mere function of the relative frequency of the verb *know*, but rather a function of its factive status (as stipulated by Hintikka and many other modal logicians) and its role in helping reasoners construct models of the premises. Future studies should investigate these sensible deductions.

One unexplored aspect of the theory is how reasoners interpret negations of mental state relations. It is challenging to stipulate a uniform semantics for negations such as these:

43a. Vira doesn't know that it's raining in Santiago.
b. Pete doesn't believe that it's raining in Santiago.
c. It's not the case that Vira knows that it's raining in Santiago.
d. It's not the case that Pete believes that it's raining in Santiago.

because they are ambiguous: (43a) could mean that Vira believes it's not raining in Santiago, or else it could mean that Vira has no belief whatsoever about Santiago's weather. In contrast, (43b) could express uncertainty on Pete's part, or else it could mean that Pete believes that it's not raining in Santiago with certainty. And both (43c) and (43d) are ambiguous in similar ways. The model-based analysis we provide above doesn't address these assertions, and empirical research has yet to study them. But previous model-theoretic treatments of negation in sentential reasoning (Khemlani et al., 2012, 2014; Orenes et al., 2014) anticipate at least three central patterns for investigation: first, negative assertions should be more difficult to process than analogous affirmative assertions (Wason, 1959, 1961). Second, the difficulty of interpreting them should interact with the number of models reasoners have to construct: negated disjunctions (which yield one model) are easier to process than negated conjunctions (which yield multiple models), while affirmative conjunctions are easier than affirmative disjunctions (Khemlani et al., 2014). Third, reasoners should discover heuristics to ease the burden of interpreting negations, e.g., they should reveal preferences and biases in how they interpret the negation in (43c).

The theory and the experiments we describe focus on people's ability to explicitly consider mental states when reasoning. But, a large body of clinical and developmental work focuses on peoples' abilities to consider mental states outside of conscious awareness – that is, their *implicit* mentalization abilities (Fonagy & Luyten, 2009; Frith & Frith, 2003; Gawronski & Bodenhausen, 2006; Kovács et al., 2010; Kulke, Johannsen, & Rakoczy, 2019; Luyten, Malcorps, Fonagy, & Ensink, 2019; Schneider, Slaughter, & Dux, 2017). This body of work likewise examines individuals' ability to automatically consider mental states even in scenarios that do not make them relevant. The present account may help explain some of the cognitive computations that underlie people's tendency to mentalize implicitly and automatically, and these patterns may help explain why individuals robustly and consistently committed omniscience errors: their consideration of mental states may have been a result of more general tendencies to mentalize information. Future research should investigate developmental and clinical ramifications of the results we describe.

Critics may wonder if the theory is falsifiable, given that it can account for both optimal and erroneous inferences. The theory does not predict *all* or *any* erroneous inference, however. For instance, no mechanism in the theory permits reasoners to make this (invalid, nonsensical) inference:

Allen knows that it's raining and windy.
Therefore, Allen knows that it's not raining.

If reasoners routinely drew conclusions such as this one, the present theory could not account for them. Likewise, the theory predicts that reasoners should rarely make inferences such as:

Jesse believes that if it's raining, it's windy.
Jesse believes that it's raining.
Therefore, Jesse knows that it's windy.

because the epistemic verb in the conclusion doesn't reflect the tags that correspond to the premises. Experiment 4 shows that such inferences are uncommon (3% of trials), but if reasoners had produced them more often, the theory could not explain such behavior. In contrast, reasoners should be very likely to draw the conclusion if the verb *believe* in the preceding example was replaced with a factive verb, such as *know* or *understand* or *realize* – and indeed they do so 77% of the time (see Table 5). Had reasoners failed to draw such conclusions, the theory would not be able to explain their reluctance.

The theory relies on the concept of factivity as a way of distinguishing epistemic tags, and so critics may worry that it leaves itself open to those who argue against the idea that certain verbs trigger factive interpretations. For instance, Hazlett (2012) observes that certain uses of *know* can be non-factive, as in:

44. Everyone knew that stress caused ulcers, before two Australian doctors in the early 80s proved that ulcers are actually caused by bacterial infection.

In (44), *know* is used to contrast a mental state against a factual matter, and so it conveys high confidence that stress causes ulcers but doesn't entail anything about its truth, and the remainder of the sentence describes why the confidence was misplaced. Whether verbs themselves are inherently factive in nature, or whether context, content, and pragmatics make them factive, is a matter of debate by epistemologists (Bricker, 2023; Dahlman & van de Weijer, 2022; Tsohatzidis, 2012). Nevertheless, as Hazlett (2012) anticipates in his analysis, *know* may presuppose the truth of its complement without entailing or guaranteeing it. In (44) above, for instance, a reasoner may presuppose that stress causes ulcers until the word "before", which serves to cancel that presupposition. Indeed, the example suggests that reasoners may construct and remove epistemic tags online as they interpret complex sentences.

Can other accounts of epistemic reasoning explain the results we describe? One recent computational account treats social and epistemic reasoning as Bayesian inference over a mental model of a rational agent (Jara-Ettinger & Rubio-Fernandez, 2021). This formal account describes how reasoners infer an individual's mental states by considering the contents of the beliefs that would cause a rational agent to communicate in the manner the reasoner observes, and it explains how people update mental state inferences from ongoing communication. Jara-Ettinger and Rubio-Fernandez posit that in everyday communication, people may rapidly analyze and deconstruct word choices to infer communicative intent, and that endogenous and exogenous communicative demands can promote rational mental state reasoning. The theory we outline can serve as a foundational layer that bridges the rational model's formal framework and imbues it with predictions about the representations and processes that occur during mental state inference. Yet the research also challenges rational models to explain systematic mistakes that reasoners make, and no other theory at present can account for the patterns outlined in Table 1.

In sum, words such as *know*, *think*, and *believe* are used to talk about the mental states an agent possesses. The model theory we describe further argues that such words trigger simulations of those mental states, which are encoded and processed to yield patterns of inference. Those simulations are modal in nature: they concern what's possible and what's impossible given the syntax, semantics, and pragmatics of the information conveyed. And the mental representation of those possibilities places demands on a capacity-limited cognitive system. Reasoning about epistemic matters is necessary for effective communication – and yet capacity-limited cognitive systems are bound to make errors. The theory we describe is the first account of mental state reasoning that explains both rational and error-prone reasoning.

Declarations of competing interest

None.

CRediT authorship contribution statement

Branden J. Bio: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Sangeet Khemlani:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization.

Appendix A. Omniscience in logic and language

A consensus in contemporary cognitive science is that humans do not reason by recourse to any symbolic logic (Khemlani, 2018; Johnson-Laird & Khemlani, 2023; Elqayam & Over, 2013; Oaksford & Chater, 2007; cf. Bringsjord & Sundar Govindarajulu, 2020). But various systems of logic, including probability logics, continue to serve as benchmarks for accurate reasoning, both for the development of psychological theory (Pfeifer & Kleiter, 2009; Pietarinen, 2003) as well as in artificial intelligence (Sutcliffe, 2017). A prominent example is the usage of epistemic logics to model valid reasoning about mental states (Bolander, 2018; van de Pol, van Rooij, & Szymanik, 2018; van Ditmarsch & Labuschagne, 2007).

Theorists developed epistemic logics to capture the modal properties of operators for knowledge and belief (Fagin, Moses, Halpern, & Vardi, 1995; Hintikka, 1962; von Wright, 1951). They argued that to express that an individual knows something – *A knows P,* or $K_A(P)$ – is to express that *P* is true in one or more situations consistent with *A's* mental state. A countable infinity of epistemic logics exist: each separate logic denotes a distinct set of axioms that describe what can and cannot follow. Here are two axioms embodied in the most frequently used epistemic logics, along with their English translations:

Axiom T:
$K_A(\mathrm{P}) \rightarrow \mathrm{P}$
(If A knows P*, then* P *is the case,* i.e.*, knowledge is factual.)*

Axiom K:
$K_A(\mathrm{P} \rightarrow \mathrm{Q}) \rightarrow (K_A(\mathrm{P}) \rightarrow K_A(\mathrm{Q}))$
(If A knows that if P then Q*, then whenever* A knows P*, then it follows that* A knows Q *too.)*

Axiom T expresses the notion that *A knows P* is true whenever *P* is true in both *A's* mental states as well as the world at large. The two axioms, and indeed, most axioms in epistemic logic, describe what can be derived from an agent's state of knowledge. No axioms describe valid deductions about the world from a set of belief states, and so as a consequence, to say that *A believes P* is to express that *P* is true in *A's* mental states but not necessarily the possible states of the world.

Critics of epistemic logic worry that it presents an implausible description of human reasoning. Axiom K above, after all, suggests that agents have immediate access to the logical consequences of their knowledge – a form of "logical omniscience" (see Stalnaker, 1991) – and early theorists such as Hintikka acknowledged this property as a discrepancy between logic and natural language (1962, p. 30–31). Nevertheless, such logics can be useful in justifying certain commonsense intuitions. Consider again problem (1) above. In epistemic logic, we might express (1) as follows:

1′. $\boldsymbol{K}_{Devon}(\mathrm{client}(Olga) \rightarrow \mathrm{student}(Olga))$.
client(Olga).

Intuitions suggest that it is a mistake to conclude that Devon knows whether or not Olga is a student, because there is no reason to believe that Devon knows she is a client. A reasoner who draws such a conclusion has made a gross error of omniscience – they presume that Devon has much more knowledge about the situation than the premises suggest. The intuition accords with all systems of epistemic logic, including the most permissive calculi, which treat this inference: $\boldsymbol{K}_{Devon}(\mathrm{student}(Olga))$ as invalid.

Appendix B. Semantic coherence manipulation check for Experiment 1

Participants across Experiments 1a-e responded in a manner that revealed systematic and careful evaluation of the premises. In both natural language and in epistemic logic, knowledge implies belief, but not vice versa. For example, if an agent *knows* that it is cloudy, then the agent believes that it is cloudy, but belief does not imply knowledge. Hence, if participants understood and processed the meanings of verbs sensibly, they should be more likely to switch epistemic verbs in their responses for factives versus non-factives. For example, consider problem (1) in the main text:

1. Devon knows that if Olga is a client, she's also a student.
   Olga is a client.

In the reported studies, omniscience errors were those in which participants responded "Devon knows that Olga is a student" or else "Devon believes that Olga is a student." Participants should be more likely to switch epistemic verbs for *know* than *believe,* i.e., they should be more likely to respond "Devon believes that Olga is a student" when the epistemic verb is *knows* and less likely to respond "Devon knows that Olga is a student" if the verb is *believe.* They did not exhibit this pattern in Experiment 1a, suggesting that participants may have evaluated the premises in a shallow manner; however, in Expeiments 1b-d, their answers revealed deliberative evaluation of the epistemic verbs, i.e., they switched from a factive ("knows") to a non-factive verb ("believes") more often than vice versa. Fig. A1 shows participants' pattern of verb switches across various versions of Experiment 1.

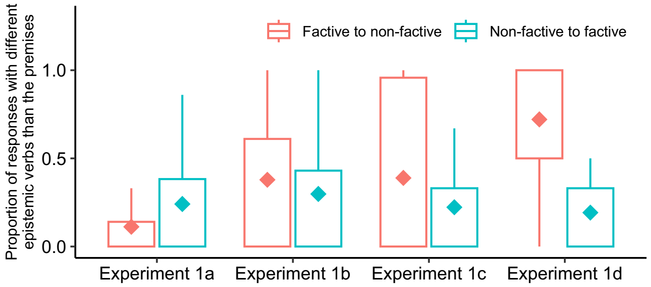

Fig. A1. Box plots of proportions of verb switches in Experiments 1a-d as a function of whether that shift was from a factive verb to a non-factive one and vice versa. Verb switches concern verb switches for those problems on which participants made omniscience errors. Light circles denote individual participants' mean proportions of omniscience errors; diamonds denote mean proportions across all participants; error bars denote 95% confidence intervals.

Data availability

Data is posted on OSF and link is provided in manuscript

References

Abbeduto, L., & Rosenberg, S. (1985). Children's knowledge of the presuppositions of know and other cognitive verbs. *Journal of Child Language, 12*, 621–641.

Apperly, I. A., Samson, D., & Humphreys, G. W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology, 45*, 190–201.

Arslan, B., Hohenberger, A., & Verbrugge, R. (2017). Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLoS One, 12*, Article e0169510.

Austin, G., Groppe, K., & Elsner, B. (2014). The reciprocal relationship between executive function and theory of mind in middle childhood: A 1-year longitudinal perspective. *Frontiers in Psychology, 5*, 1–11.

Baddeley, A. D., Hitch, G., & Allen, R. (2021). *A multicomponent model of working memory* (pp. 10–43). Working memory: State of the science.

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.

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

Barrouillet, P., Grosset, N., & Lecas, J. F. (2000). Conditional reasoning by mental models: Chronometric and developmental evidence. *Cognition, 75*, 237–266.

Bell, V., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science, 22*, 25–51.

Bendaña, J., & Mandelbaum, E. (2021). The fragmentation of belief. In C. Borgoni, D. Kindermann, & A. Onofri (Eds.), *The fragmented mind*. Oxford: Oxford University Press.

Bianco, F., Lombardi, E., Lecce, S., Marchetti, A., Massaro, D., Valle, A., & Castelli, I. (2021). Supporting children's second-order recursive thinking and advanced ToM abilities: A training study. *Journal of Cognition and Development, 22*, 561–584.

Bigelow, E. J., McCoy, J. P., & Ullman, T. D. (2023). Non-commitment in mental imagery. *Cognition, 238*, 105498.

Bio, B., & Khemlani, S. (2023). Omniscience errors in mental state reasoning. In M. Goldwater, F. Anggoro, B. Hayes, & D. Ong (Eds.), *Proceedings of the 45th annual conference of the cognitive science society*. Cognitive Science Society: Austin, TX.

Bio, B. J., Guterstam, A., Pinsk, M., Wilterson, A. I., & Graziano, M. S. (2022). Right temporoparietal junction encodes inferred visual knowledge of others. *Neuropsychologia, 171*, Article 108243.

Bio, B. J., Webb, T. W., & Graziano, M. S. (2018). Projecting one's own spatial bias onto others during a theory-of-mind task. *Proceedings of the National Academy of Sciences, 115*(7), e1684–e1689.

Birch, S. A., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences, 8*, 255–260.

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*.

Bolander, T. (2018). Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In H. van Ditmarsch, & G. Sandu (Eds.), *Jaakko Hintikka on knowledge and game-theoretical semantics* (pp. 207–236). Cham: Springer.

Bouchacourt, F., & Buschman, T. J. (2019). A flexible model of working memory. *Neuron, 103*, 147–160.

Bricker, A. M. (2023). Knowledge as a (non-factive) mental state. *Erkenntnis*, 1–22.

Bringsjord, S., & Sundar Govindarajulu, N. (2020). Rectifying the mischaracterization of logic by mental model theorists. *Cognitive Science, 44*, Article e12898.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology, 50*, 159–193.

Byrne, R. M. (2005). *The rational imagination*. MIT Press.

Byrne, R. M. (2017). Counterfactual thinking: From logic to morality. *Current Directions in Psychological Science, 26*, 314–322.

Byrne, R. M., Evans, J. S. B., & Newstead, S. E. (2019). *Human reasoning: The psychology of deduction*. Psychology Press.

Byrne, R. M., Evans, J. S. B., & Newstead, S. E. (1993). *Human reasoning: The psychology of deduction*. Psychology Press.

Carey, S., Leahy, B., Redshaw, J., & Suddendorf, T. (2020). Could it be so? The cognitive science of possibility. *Trends in Cognitive Sciences, 24*, 3–4.

Cohen, S. (2002). Basic knowledge and the problem of easy knowledge. *Philosophy and Phenomenological Research, 65*, 309–329.

Cortes, R. A., Weinberger, A. B., Colaizzi, G. A., Porter, G. F., Dyke, E. L., Keaton, H. O., & Green, A. E. (2021). What makes mental modeling difficult? Normative data for the multidimensional relational reasoning task. *Frontiers in Psychology, 12*, Article 668256.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19*, 274–282.

Dahlman, R. C., & van de Weijer, J. (2022). Cognitive factive verbs across languages. *Language Sciences, 90*, Article 101458.

De Brigard, F. (2023). Counterfactual thinking. In *The Palgrave encyclopedia of the possible* (pp. 243–250). Cham: Springer International Publishing.

Diamond, A., & Kirkham, N. (2005). Not quite as grown-up as we like to think: Parallels between cognition in childhood and adulthood. *Psychological Science, 16*, 291–297.

van Ditmarsch, H., & Labuschagne, W. (2007). My beliefs about your beliefs: A case study in theory of mind and epistemic logic. *Synthese, 155*, 191–209.

Elga, A., & Rayo, A. (2022). Fragmentation and logical omniscience. *Noûs, 56*, 716–741.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking & Reasoning, 19*, 249–265.

Everett, D. L. (2012). What does Pirahã grammar have to teach us about human language and the mind? *Wiley Interdisciplinary Reviews: Cognitive Science, 3*, 555–563.

Fagin, R., Moses, Y., Halpern, J. Y., & Vardi, M. Y. (1995). Knowledge-based programs. In *Proceedings of the 14th Annual ACM Symposium on Principles of Distributed Computing*.

Fonagy, P., & Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Development and Psychopathology, 21*, 1355–1381.

Frank, M. C., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science, 336*, 998.

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

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological bulletin, 132*(5), 692.

Gerstenberg, T. (2024). *Counterfactual simulation in causal cognition*. Manuscript in press at *Trends in Cognitive Sciences*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review, 128*, 936.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565–610.

Harner, H., & Khemlani, S. (2022). Reasoning about *want*. *Cognitive Science, 46*, Article e13170.

Hazlett, A. (2012). Factive presupposition and the truth condition on knowledge. *Acta Analytica, 27*(4), 461–478.

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1606.

Hintikka, K. J. J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition, 35*, 2075–2087.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). Quantitative mental state attributions in language understanding. *Science Advances, 7*, Article eabj0970.

Jeffrey, R. C. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York: McGraw-Hill.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*, 646.

Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99*, 418.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Erlbaum.

Johnson-Laird, P. N., Byrne, R. M. J., & Khemlani, S. (2024). Models of possibilities instead of logic as the basis of human reasoning. *Minds and Machines, 34*, 19.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111*, 640.

Johnson-Laird, P. N., & Khemlani, S. (2023). Mental models and the algorithms of deduction. In R. Sun (Ed.), *Cambridge handbook of computational cognitive sciences*.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences, 19*, 201–214.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science, 288*, 531–532.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review, 106*, 62.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition, 193*, 103950.

Johnson-Laird, P. N., & Ragni, M. (2024). Reasoning about possibilities: Modal logics, possible worlds, and mental models. *Manuscript in press at Psychonomic Bulletin & Review*.

Juhos, C., Quelhas, A. C., & Johnson-Laird, P. N. (2012). Temporal and spatial relations in sentential reasoning. *Cognition, 122*, 393–404.

Kelly, L. J., & Khemlani, S. (2023). Iconicity bias and duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Kelly, L. J., Khemlani, S., & Johnson-Laird, P. N. (2020). Reasoning about durations. *Journal of Cognitive Neuroscience, 32*, 2103–2116.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25–41.

Khemlani, S. (2018). Reasoning. In S. Thompson-Schill (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience*. Wiley & Sons.

Khemlani, S. (2021). Epistemic verbs produce spatial models. In T. Fitch, C. Lamm, H. Leder, & K. Tessmar (Eds.), *Proceedings of the 43rd annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Khemlani, S., Byrne, R. M., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science, 42*, 1887–1924.

Khemlani, S., & Johnson-Laird, P. N. (2017). Illusions in reasoning. *Minds and Machines, 27*, 11–35.

Khemlani, S., & Johnson-Laird, P. N. (2022). Reasoning about properties: A computational theory. *Psychological Review, 129*, 289.

Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *proceedings of the national academy of sciences, 110*(42), 16766–16771.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology, 24*, 541–559.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2014). The negations of conjunctions, conditionals, and disjunctions. *Acta Psychologica, 151*, 1–7.

Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & Cognition, 46*.

Khemlani, S. S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology, 64*, 2276–2288.

Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & cognition, 30*(3), 363–371.

Korman, J., & Khemlani, S. (2020). Explanatory completeness. *Acta Psychologica, 209*, Article 103139.

Kouklari, E. C., Thompson, T., Monks, C. P., & Tsermentseli, S. (2017). Hot and cool executive function and its relation to theory of mind in children with and without autism spectrum disorder. *Journal of Cognition and Development, 18*(4), 399–418.

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

Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit theory of mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PLoS One, 14*, Article e0213772.

Kumar, A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review, 28*, 40–80.

Logie, R., Camos, V., & Cowan, N. (Eds.). (2020). *Working memory: The state of the science*. Oxford, UK: Oxford University Press.

Luyten, P., Malcorps, S., Fonagy, P., & Ensink, K. (2019). Assessment of mentalizing. *Handbook of mentalizing in mental health practice, 2*, 37–62.

Margoni, F., & Brown, T. (2023). Jurors use mental state information to assess breach in negligence cases. *Cognition, 236*, Article 105442.

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review, 22*, 5–36.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Harvard University Press.

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin, 135*, 749.

Nazlidou, E.-I., Moraitou, D., Natsopoulos, D., Papaliagkas, V., Masoura, E., & Papantoniou, G. (2018). Inefficient understanding of non-factive mental verbs with social aspect in adults: Comparison to cognitive factive verb processing. *Neuropsychiatric Disease and Treatment, 14*, 2617–2631.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Orenes, I., Beltrán, D., & Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language, 74*, 36–45.

Orenes, I., & Johnson-Laird, P. N. (2012). Logic, models, and paradoxical inferences. *Mind & Language, 27*, 357–377.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making, 5*, 411–419.

Peirce, C. S. (1931–1958). In C. Hartshorne, P. Weiss, & A. Burks (Eds.), *Collected papers of Charles Sanders Peirce*. Harvard University Press.

Peloquin, C. (2021). *Representing minds representing minds: An examination of the association between recursive mental state attributions and executive processes*. Doctoral dissertation. Victoria University of Wellington.

Pfeifer, N., & Kleiter, G. D. (2009). Mental probability logic. *Behavioral and Brain Sciences, 32*, 98–99.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences, 23*, 1026–1040.

Pietarinen, A. V. (2003). What do epistemic logic and cognitive science have to do with each other? *Cognitive Systems Research, 4*, 169–190.

van de Pol, I., van Rooij, I., & Szymanik, J. (2018). Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. *Journal of Logic, Language and Information, 27*.

Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology, 63*, 1716–1739.

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2019). The analytic truth and falsity of disjunctions. *Cognitive Science, 43*(9), e12739.

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief, 2*. New York: Random House.

Ragni, M., Khemlani, S., & Johnson-Laird, P. N. (2014). The evaluation of the consistency of quantified assertions. *Memory & Cognition, 42*, 53–66.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review, 120*, 561.

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From knowing to remembering: The semantic–episodic distinction. *Trends in Cognitive Sciences, 23*, 1041–1057.

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003). "I know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology, 7*, 38–65.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and U.S. preschoolers. *Psychological Science, 17*, 74–81.

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science, 23*, 842–847.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic theory of mind processing in adults. *Cognition, 162*, 27–31.

Schroyens, W. J., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking & Reasoning, 7*, 121–172.

Shah, A. K., & LaForest, M. (2022). Knowledge about others reduces one's own sense of anonymity. *Nature, 603*, 297–301.

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition, 14*, 301–321.

Shetreet, E., Alexander, E. J., Romoli, J., Chierchia, G., & Kuperberg, G. (2019). What we know about knowing: Presuppositions generated by factive verbs influence downstream neural processing. *Cognition, 184*, 96–106.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*, 171–177.

Stalnaker, R. (1991). The problem of logical omniscience. *I. Synthese*, 425–440.

Surtees, A. D., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology, 30*, 75–86.

Sutcliffe, G. (2017). The TPTP problem library and associated infrastructure. *Journal of Automated Reasoning, 59*, 483–502.

Tsohatzidis, S. L. (2012). How to forget that "know" is factive. *Acta Analytica, 27*, 449–459.

Tullis, J. G., & Feder, B. (2023). The "curse of knowledge" when predicting others' knowledge. *Memory & Cognition, 51*, 1214–1234.

Unsworth, N., & Robison, M. K. (2020). Working memory capacity and sustained attention: A cognitive-energetic perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 77.

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experiment Psychology, 11*, 92–107.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology, 52*, 133–142.

Wason, P. C., & Johnson-Laird, P. (1972). *Psychology of reasoning: Form and content.* Harvard University Press.

Wellman, H. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology, 15*, 728–755.

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

Woo, B., Chisholm, G., & Spelke, E. (2024). Do toddlers reason about other people's experiences of objects? A limit to early mental state reasoning. *Cognition, 246*, Article 105760.

von Wright, G. H. (1951). *An essay in modal logic*. Amsterdam: North Holland.

Zhao, W. J., Richie, R., & Bhatia, S. (2022). Process and content in decisions from memory. *Psychological Review, 129*, 73.

Further-reading

Johnson-Laird, P. N., Byrne, R. M., & Khemlani, S. (2023). Human verifications: Computable with truth values outside logic. *Proceedings of the National Academy of Sciences, 120*, Article e2310488120.