COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13309

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13309

Do Large Language Models Know What Humans Know?

Sean Trott,* Cameron Jones,* Tyler Chang, James Michaelov, Benjamin Bergen

Department of Cognitive Science, University of California San Diego

Received 6 January 2023; received in revised form 2 June 2023; accepted 12 June 2023

Abstract

Humans can attribute beliefs to others. However, it is unknown to what extent this ability results from an innate biological endowment or from experience accrued through child development, particularly exposure to language describing others' mental states. We test the viability of the language exposure hypothesis by assessing whether models exposed to large quantities of human language display sensitivity to the implied knowledge states of characters in written passages. In pre-registered analyses, we present a linguistic version of the False Belief Task to both human participants and a large language model, GPT-3. Both are sensitive to others' beliefs, but while the language model significantly exceeds chance behavior, it does not perform as well as the humans nor does it explain the full extent of their behavior—despite being exposed to more language than a human would in a lifetime. This suggests that while statistical learning from language exposure may in part explain how humans develop the ability to reason about the mental states of others, other mechanisms are also responsible.

Keywords: Large language models; Language; False Belief Task; Belief attribution

1. Introduction

Humans reason about the beliefs of others, even when these beliefs diverge from their own. The capacity to understand that others' beliefs can differ from ours—and from the

*These authors contributed equally to this work.
Correspondence should be sent to Cameron Jones, Department of Cognitive Science, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. E-mail: cameron@ucsd.edu

truth—appears critical for human social cognition (Fairchild & Papafragou, 2021; Leslie, 2001). Yet despite consensus on the importance of belief attribution, there remains considerable debate about its evolutionary (Krupenye & Call, 2019a; Premack & Woodruff, 1978) and developmental (Bedny, Pascual-Leone, & Saxe, 2009; de Villiers & de Villiers, 2014) origins. Specifically, how much of this ability results from an innate, biologically evolved adaptation (Bedny et al., 2009), and how much is assembled from experience (Hughes et al., 2005)?

The answers to these questions may also tell us what kinds of biological and artificial entities can be expected to display the ability to reason about other agents' beliefs—and perhaps display evidence of social cognition more generally. The ability to represent others' beliefs has sometimes been linked to a broader constellation of abilities called theory of mind (Apperly, 2012; Premack & Woodruff, 1978). However, the theoretical, convergent, and predictive validity of such a construct has been widely questioned (Gernsbacher & Yergeau, 2019; Gough, 2021; Hayward & Homer, 2017). We focus more narrowly here on belief attribution; whether or not it is a component of a broader capacity, the ability to attribute beliefs to others is likely to be crucial to social cognition and worthy of careful analysis in its own right. We return in our Discussion to the relevance of our results to broader debates about how belief attribution relates to other capacities.

A leading experience-based view of the origins of belief attribution proposes that our ability to represent others' beliefs is built in part from exposure to language (de Villiers & de Villiers, 2014). Children develop an understanding that others have different mental states from verbs like "know" and "believe" (Brown, Donelan-McCall, & Dunn, 1996), the structure of conversation (Harris, 2005), and certain syntactic structures, like sentential complements (e.g., "Mary thought that Fred went to the movies"; Hale & Tager-Flusberg, 2003).

However, current evidence does not address the question of *the extent* to which linguistic input alone can account for the ability to reason about beliefs. Can human-level sensitivity to the beliefs of others emerge out of exposure to linguistic input by itself, or does it depend on linking that input to a distinct (possibly innate) mechanism or to non-linguistic experiences or representations? Answering these questions would require a measure of sensitivity to the beliefs of others, as well as an operationalization of what kinds of behavior can be acquired through exposure to language alone.

Only recently has constructing such a measure become tractable, with the advent of large language models (LLMs). Language models learn to assign probabilities to word sequences based on statistical patterns in the way that words are distributed in language. While early n-gram models simply learn transition probabilities between one sequence of words and the next, modern language models use neural networks to represent words in a multidimensional meaning space, allowing them to generalize to sequences they have never observed before (Jurafsky & Martin, 2019). Additionally, they contain attention mechanisms that allow them to relate words in the input stream to one another and represent each word differently depending on its context (Vaswani et al., 2017). Modern LLMs are neural language models with billions of parameters trained on corpora of hundreds of billions of words. We ask whether LLMs' considerable sensitivity to distributional patterns allows them to systematically assign higher probabilities to word sequences that describe plausible belief attribution—a behavior which is thought to result from reasoning about the beliefs of others in humans.

As others have noted (Bender & Koller, 2020), the training regime for LLMs does not include social interaction, experience in a physical environment, or even the notion of communicative intent.[1] Most relevant to the current question, their network architecture is also not pre-coded with any conception of social agents or the ability to reason about and attribute beliefs to others. And yet, LLMs have recently been shown to display a range of surprising behaviors consistent with the acquisition of linguistic structure (Linzen & Baroni, 2021; Manning, Clark, Hewitt, Khandelwal, & Levy, 2020; Sinclair, Jumelet, Zuidema, & Fernández, 2022) and arguably certain aspects of linguistically conveyed meaning (Abdou et al., 2021; Li, Nye, & Andreas, 2021). LLMs have also been the subject of recent public discussion (Johnson, 2022), including speculation that they can acquire something akin to theory of mind. They thus serve as useful baselines for what kinds of behavior can be produced merely by exposure to distributional statistics of language in general, and for belief attribution in particular. Specifically, if LLMs display sensitivity to implied belief states, it may undermine the claim that other mechanisms (i.e., either an innate biological endowment or non-linguistic sources of experience) are *necessary* for the development of this capacity.

In two pre-registered analyses, we investigated whether GPT-3 (Brown et al., 2020), a state-of-the-art LLM, displayed sensitivity to implied belief states using the widely used False Belief Task (Wimmer & Perner, 1983). It is worth acknowledging from the outset that the False Belief Task has been criticized on several grounds (Bloom & German, 2000), both because it is too narrow (it does not measure participants' abilities to reason about other mental states such as emotions and intentions) and too broad (successful performance likely involves capacities beyond reasoning about beliefs, such as executive function). Our study is therefore limited in what it can say about LLMs' sensitivity to other mental states. Moreover, low performance by either human or LLM participants could be due to lacking other necessary capacities beyond belief attribution itself. Nonetheless, the False Belief Task remains a key and extensively used instrument for assessing the capacity to reason about beliefs in humans (Bradford, Brunsdon, & Ferguson, 2020; Fairchild & Papafragou, 2021; Pluta et al., 2021; Xie, Cheung, Shen, & Wang, 2018) and other animals (Krupenye & Call, 2019a; Premack & Woodruff, 1978) as well as the neural underpinnings of this capacity (Schneider, Slaughter, Becker, & Dux, 2014). It also has the advantage of being implementable using purely linguistic stimuli, which makes it amenable to comparison between humans and LLMs. It is important to highlight that human false belief accuracy is rarely perfect, so we do not compare LLMs to an idealized perfect human participant. Instead, we elicit data from both LLM and human participants. In addition to analyzing the responses from each group, we also quantify the extent to which human responses can be predicted by LLM responses.

Our implementation of the False Belief Task involves written text passages in English. We generated novel False Belief Task stimuli, to ensure that they could not have appeared in GPT-3's training data and presented the same stimuli to humans and GPT-3. In each passage, a character places an object in a Start location, and the object is subsequently moved to an End location. The key manipulation was the knowledge state of the main character. In the False Belief condition, the character is not present when the object is moved and is thus unlikely to know that it has changed location; in the True Belief condition, the character is present and is thus more likely to know that it is in a new location. To control for other factors that might

impact belief judgments, we orthogonally counterbalanced whether the first mention and most recent mention of a location was the Start or End location; we also ensured that the start and end locations were mentioned an equal number of times in each passage. Humans and the LLM then completed a cue sentence indicating the character's belief about the object location. This knowledge cue was either Explicit ("Sean thinks the book is in the…") or Implicit ("Sean goes to get the book from the…"). The correct completion (i.e., the one consistent with the beliefs of the character) was the End location on True Belief trials and the Start location on False Belief trials. We measured whether participants responded with the Start or End location and the relative probability that GPT-3 assigned to the Start location (the log-odds of Start vs. End).

We ask two key questions. First, are LLMs sensitive to false belief? That is, is GPT-3 sufficiently sensitive to information in a preceding sequence (describing a character's beliefs) that it assigns a higher probability to subsequent sequences which describe behavior consistent with those beliefs versus subsequent sequences describing inconsistent behavior? If so, then biological and artificial agents could in principle develop behavior consistent with sensitivity to false beliefs from input-driven mechanisms alone, such as exposure to language (de Villiers & de Villiers, 2014). Notably, this empirical result could be used either to support claims that LLMs implicitly represent the belief states of others—as success at the False Belief Task is often interpreted for infants and non-human animals (Krupenye & Call, 2019a)—or as evidence that the False Belief Task is not a valid measure of mentalizing ability; this issue is explored in greater detail in the Discussion.

The second question is whether LLMs *fully* explain human behavior in the False Belief Task. If so, this would show that language exposure is not only a viable mechanism in general but that it may in fact be *sufficient* to explain how humans in particular come to display sensitivity to the belief states of others. Importantly, this would undermine claims that other non-linguistic resources are necessary to account for the human ability to reason about the beliefs of others. If LLMs do not fully explain human behavior, however, we infer that humans rely on some other mechanism not available to the LLM, such as an innate capacity or experience with more than just language.

All experiments and analyses were pre-registered on the Open Source Framework. The pre-registered analysis of LLM sensitivity can be found here: https://osf.io/agqwv. The pre-registration for the human experiment and analysis can be found here: https://osf.io/zp6q8.

2. Method

2.1. False Belief passages

We constructed 12 template passages (items) that conformed to the standard False Belief Task structure (Wimmer & Perner, 1983). In each, a main character puts an object in a Start location and a second character moves the object to an End location. The last sentence of each passage states or implies that the main character believes the object is in some (omitted) location (e.g., "Sean thinks the book is in the...."). We created 16 versions of each item

(192 passages) that varied across four dimensions. The primary dimension was knowledge state: whether the main character knows (True Belief) or does not know (False belief) that the object has changed location; this was manipulated by the main character being present or not when the second character moved the object. We also manipulated whether the first mention and most recent mention of a location is the Start or End location; and knowledge cue: whether the main character's belief is stated implicitly (e.g., "goes to get the book from the…") or explicitly (e.g., "thinks the book is in the…"). Each location was mentioned twice in each passage. In the example passages, below, the first mention is to the Start location and the recent mention is to the End location.

> **True Belief Passage**: "Sean is reading a book. When he is done, he puts the book in the box and picks up a sweater from the basket. Then, Anna comes into the room. Sean watches Anna move the book from the box to the basket. Sean leaves to get something to eat in the kitchen. Sean comes back into the room and wants to read more of his book."
>
> **False Belief Passage**: "Sean is reading a book. When he is done, he puts the book in the box and picks up a sweater from the basket. Then, Anna comes into the room. Sean leaves to get something to eat in the kitchen. While he is away, Anna moves the book from the box to the basket. Sean comes back into the room and wants to read more of his book."
>
> **Implicit Cue**: "Sean goes to get the book from the...."
>
> **Explicit Cue**: "Sean thinks the book is in the...."

2.2. GPT-3 log-odds

We used GPT-3 *text-davinci-002* to estimate the distributional likelihood of different passage completions. GPT-3 *text-davinci-002* is based on GPT-3 *davinci* (Brown et al., 2020), a 175B unidirectional LLM trained on hundreds of billions of tokens of text from the web, books, and Wikipedia. GPT-3 *text-davinci-002* (hereafter GPT-3) is additionally fine-tuned on requests to follow instructions and performs better on a variety of tasks than the original GPT-3 *davinci* (OpenAI, 2023b). We elicited from GPT-3 the log probability of each possible location (Start vs. End) at the end of each passage version, equal to the log probabilities of those locations in a free-response completion. Where a location comprised multiple tokens we summed the log probabilities. We accessed GPT-3 through the OpenAI API. Using the log-odds ratio, $\log(p(\text{Start})) - \log(p(\text{End}))$, higher values indicate larger relative probabilities of the Start location. Each passage version was presented to GPT-3 independently, and the model was not updated during inference so it did not learn across trials.

2.3. Human participant responses

A total number of 1156 participants from Amazon's Mechanical Turk platform were paid $1 for their time. Each read a single passage (except the final sentence) at their own pace. On a new page, they were asked to complete the final sentence of the passage by entering a single

word in a free-response text input. Participants then completed two free-response attention check questions that asked for the true location of the object at the start and the end of the passage. Each participant completed only one trial to prevent them from learning across the experiment, analogously to GPT-3, which saw each passage individually and could not learn across trials.

We preprocessed responses by lowercasing and removing punctuation, stopwords, and trailing whitespace. We excluded participants who were non-native English speakers (13), answered ≥ 1 attention check incorrectly (513), or answered the sentence completion with a word that was not the start or end location (17), retaining 613 trials. While this exclusion rate is unusually high, 75% of incorrect attention check responses were neither the start nor end location, indicating inattention. We implemented a parallel check for GPT-3, which responded correctly to both attention check questions on 86% of items. In our Supporting Information, we report additional analyses on incorrect responses and excluded data. After exclusions, the number of trials per item ranged from 42 to 60, and there were 313 False Belief trials and 317 True Belief trials. These pre-registered exclusion criteria reduced the likelihood that bots (Webb & Tangney, 2022) as well as participants who did not successfully comprehend the passage for any reason were included in the human data.

All research was approved by the organization's Institutional Review Board.

3. Results

3.1. Analysis of large language model behavior

In a pre-registered analysis, nested model comparisons determined whether GPT-3 log-odds changed as a function of factors such as knowledge state (False Belief vs. True Belief). We constructed a linear mixed effects model with log-odds as a dependent variable; fixed effects of knowledge state, knowledge cue, first mention, and recent mention; along with by-item random slopes for the effect of knowledge state (and random intercepts for items). This full model exhibited better fit than a model excluding only knowledge state [$\chi^2(1) = 18.6$, $p < .001$], but still preserving the other covariates (e.g., first mention, recent mention, and knowledge cue). Log-odds were lower in the True Belief condition, reflecting the correct prediction that characters should be more likely to look in the End location if they are aware that the object was moved (see Fig. 1). Critically, this main effect of knowledge state indicates that GPT-3 is sensitive to the manipulation of a character's beliefs about where an object is located. The model's raw accuracy when predicting the most probable out of the Start and End locations was 74.5%.

Additionally, the linear mixed effects model was further improved by an interaction between knowledge state and knowledge cue (Explicit vs. Implicit) [$\chi^2(1) = 20.6$, $p < 0.001$]. The effect of knowledge state was stronger in the Implicit condition [$\beta = -2.57$, $SE = 0.548$]; however, a main effect of knowledge state was found in both the Explicit [$\chi^2(1) = 13.3$, $p < .001$] and Implicit [$\chi^2(1) = 18.8$, $p < .001$] conditions. Recent mention was not a significant predictor of log-odds. However, Start completions were more

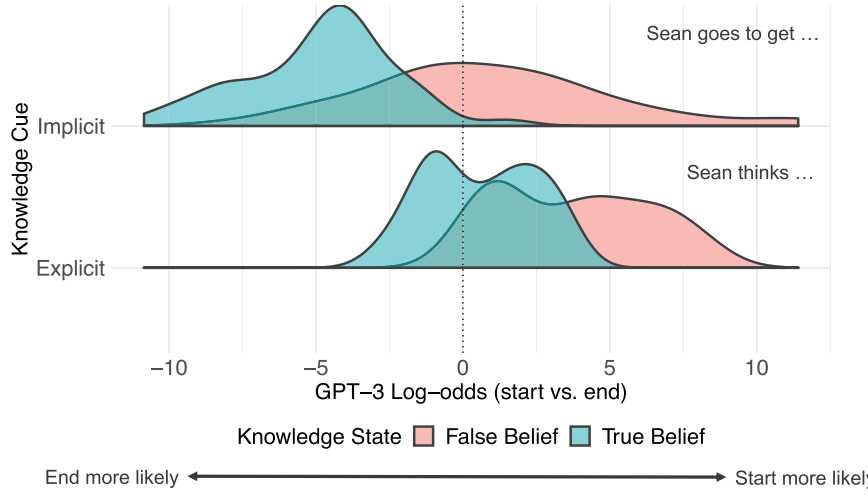

Fig. 1. GPT-3 log-odds of Start versus End location was higher (i.e., Start was relatively more likely) in the False Belief than True Belief condition [$\chi^2(1) = 18.6$, $p < .001$]. This suggests that GPT-3's predictions are sensitive to the character's implied belief state: the character is unaware that the object has moved to the End location if they did not observe it being moved. This effect was observed both when the knowledge cue was Implicit ("Sean goes to get the book from the…") and Explicit ("Sean thinks the book is in the…"); however, the effect was strengthened in the Implicit condition [$\chi^2(1) = 20.6$, $p < 0.001$].

likely when the Start location was mentioned first [$\beta = 1.32$, $SE = 0.274$, $p < 0.001$]. There was also a main effect of knowledge cue [$\beta = -2.93$, $SE = 0.388$, $p < .001$] (see Fig. 1). GPT-3 was biased towards the End location (i.e., the true location of the object) in the Implicit condition and towards the Start location in the Explicit condition. Concretely, GPT-3 predicts that explicit cues to belief state (e.g., "Sean thinks that the book is in the…" vs. "Sean goes to get the book from the…") correlate with false beliefs, demonstrating that this may be learnable from the statistics of language.

3.2. *Analysis of human responses*

Our second critical pre-registered question was whether knowledge state continued to explain human behavior even accounting for the log-odds obtained from GPT-3. We first constructed a base model predicting whether or not human participants responded with the Start location. Note that the Start location would be the correct response (i.e., congruent with knowledge states) in the False Belief condition, and the End location would be the correct response in the True Belief condition. The base model contained fixed effects of log-odds (i.e., from GPT-3), knowledge cue, first mention, and recent mention, along with by-item random slopes for the effect of knowledge state (and by-item random intercepts). Critically, this base model was significantly improved by adding knowledge state as a predictor [$\chi^2(1) = 30.4$, $p < 0.001$]. This result implies that human responses are influenced by knowledge state in a way that is not captured by GPT-3. That is, GPT-3 cannot fully account for human sensitivity to knowledge states. This is

Fig. 2. Both human participants and GPT-3 were more likely to say that a character believed an object was in the Start location when the character had not observed the object being moved to the End location (False Belief). This effect was stronger for humans than for GPT-3, and there was a marginal effect of knowledge state (True vs. False Belief) in humans that could not be accounted for by GPT-3 predictions [$\chi^2(1) = 30.4$, $p < 0.001$].

highlighted by the contrast in Fig. 2: while both human participants and GPT-3 were sensitive to knowledge state, humans displayed a much stronger effect across conditions. Additionally, mean accuracy among retained human participants (82.7%) was also higher than GPT-3 accuracy (74.5%), providing further evidence of a performance gap.

In order to test whether the high exclusion rate introduced by our attention check questions had an impact on results, we performed an exploratory analysis on all human response data. Accuracy before exclusions (including those who failed the attention checks and provided responses to neither the start or end locations) was 55.8%. After excluding responses that were neither of the start and end locations (23%), accuracy was 73.1%. It is noteworthy that this estimate of human accuracy is lower than GPT-3's performance. However, this analysis was not pre-registered. Moreover, given existing concerns about data quality on the Mechanical Turk platform (Webb & Tangney, 2022), and the fact that performance on the FB task by neurotypical adults is often assumed to be at ceiling (Dodell-Feder, Lincoln, Coulson, & Hooker, 2013), the retained data likely provide a better estimate of attentive human participant performance.

3.3. Analysis of GPT-3 token predictions

The pre-registered tasks for humans and the LLM were slightly different—humans filled in a single predicted word while we calculated the relative surprisal to two different words presented to GPT-3. To investigate whether differences in performance could be chalked up to this difference in method, we conducted an additional, exploratory analysis, in which we elicited token predictions from GPT-3. Specifically, GPT-3 was presented with the

original passage ending with the critical sentence (e.g., "Sean goes to get the book from the"), then asked to predict the upcoming word. We sampled the word (e.g., "box") with the top probability. We automatically tagged each response as correct or incorrect by checking whether GPT-3's response corresponded to the character's likely belief state about the object. That is, in the True Belief condition, the correct response would be the *End* location of the object; in the False Belief condition, the correct response would be the *Start* location of the object. When computing GPT-3 accuracy, this token generation method only differed from the pre-registered method in that GPT-3's prediction was no longer restricted to the Start or End location. Using the token generation method did not qualitatively change the results. As reported above, when assessing accuracy with the pre-registered method—relative probability assigned to start versus end locations—GPT-3 performed at 74.5% accuracy. When assessing accuracy using the more human-comparable procedure, token generation, GPT-3 performed at 73.4% accuracy, still well above chance yet now slightly farther below human accuracy.

In order to test what features of LLMs permit them to display the behaviors described above, we also tested a number of different GPT-3 models ranging in size from *ada* ($\sim$ 350M parameters) to *davinci* ($\sim$175B parameters). For each model size, we tested a *base* version, pre-trained on a large corpus of text, and a *text* version, which had additionally been fine-tuned by OpenAI using responses to human instructions. The largest fine-tuned model was *text-davinci-002*, which was the same model we used in the pre-registered analysis described above.

As expected, the largest models (*davinci*, and updated variant *text-002-davinci*) exhibited the most successful performance. The former answered correctly on 60.4% of items, while the latter answered correctly on 73.4% of items. The model with the worst performance was *text-001-ada*, which was also the smallest model (see also Fig. 3). The smallest and lowest-performing models did not exceed chance performance, which emphasizes the need for large, powerful models to succeed at this task as well as the potential, as models continue to increase in size, for LLM improvement.

4. Discussion

We asked whether exposure to linguistic input alone could account for human sensitivity to knowledge states. We found that GPT-3's predictions were sensitive to a character's implied knowledge states. When assessing accuracy with the relative probability assigned to start versus end locations, GPT-3 performed at approximately 74.5% accuracy. When assessing accuracy using token generation, the best GPT-3 model performed at 73.4% accuracy. This demonstrates that exposure to linguistic input alone can in principle account for *some* sensitivity to false belief.

However, GPT-3 was less sensitive than humans (who displayed 82.7% accuracy—see also probabilities in Fig. 2). Most critically, human behavior was not explained fully by that of GPT-3 in a statistical model. This entails that the capacities underlying human behavior in this False Belief task cannot be explained purely by exposure to language statistics—at least insofar as those statistics are reflected in GPT-3.

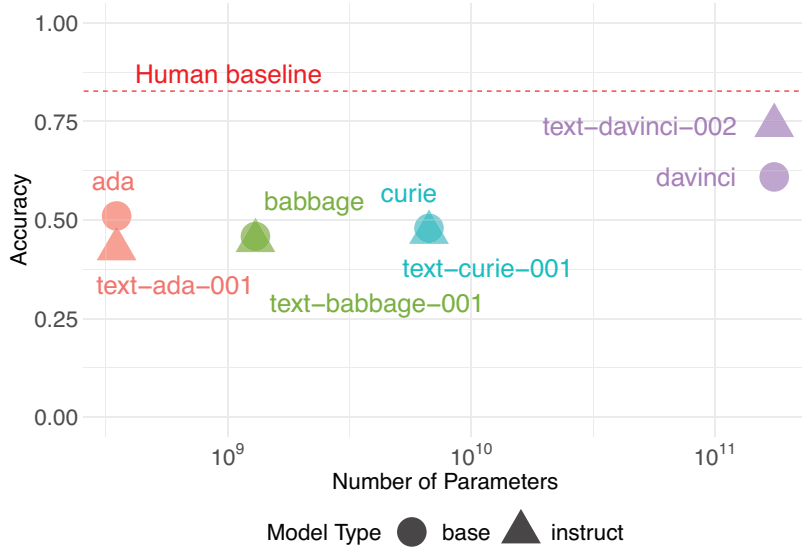

Fig. 3. A number of GPT-3 models varying in size were presented with each passage and asked to complete the critical sentence, as human participants did. For each model size, we tested a pre-trained base model (◯) and a version fine-tuned by OpenAI to follow text instructions (△). In the True Belief condition, a correct (i.e., knowledge-congruent) response corresponded to the End location of the object; in the False Belief condition, a correct (i.e., knowledge-congruent) response corresponded to the Start location of the object. The dotted red line represents human accuracy on the task (82.7%); *text-davinci-002*—the largest fine-tuned model—came the closest to approaching human behavior, with an accuracy of 73.4%.

4.1. Do LLMs attribute beliefs?

With increased academic and public attention on how humanlike LLMs are, it is worth considering what these findings imply about the cognitive capacities of AIs. First, it is important to note that—as mentioned in the Introduction—the False Belief Task is designed to measure a specific capacity: the ability to reason about the belief states of others and use that information to make predictions about their behavior. The current work cannot address whether LLMs display other purported aspects of theory of mind, including inferring implicit emotional states and reasoning about the intended interpretation of an utterance; this issue is explored at greater length later in the Discussion.

On the specific question of whether LLMs are sensitive to belief states, the evidence presented here is mixed. State-of-the-art models display sensitivity to the beliefs of others in a False Belief Task, a behavior that would have been unthinkable a few years ago from a statistical learner and indeed is only shown by the largest, highest performing LLMs. Yet they still do not achieve human-level performance. There are several possible interpretations of this result, each of which carries significant consequences for the broader debate about the nature and origins of the ability to attribute belief states.

4.1.1. Competing interpretations

One interpretation, which we call the **duck test** position, is that we should ascribe cognitive properties to agents based on observable behavioral criteria: "if it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck"; this view is roughly analogous to what is sometimes called the *superficial* view (Schwitzgebel, 2013; Shevlin, 2022) or the *intentional stance* (Dennett, 1978). False Belief Task performance has been used to support claims that infants (Baillargeon, Scott, & He, 2010) and non-human animals (Krupenye, Kano, Hirata, Call, & Tomasello, 2016) can represent the beliefs of others. The duck test view argues that the same evidence should be equally persuasive in the case of intelligent artificial agents. Although LLMs do not show the same sensitivity to belief states that humans do, this could suggest that LLMs display a less developed form of the ability to attribute belief states that is nonetheless qualitatively similar to that of humans. Under this interpretation, the ability to attribute and reason about belief states lies on a continuum, and further developments in LLMs (e.g., larger models, more training data) could lead them to more closely approximate human behavior (Kaplan et al., 2020).

The duck test view has two important implications for debates around language models and false belief sensitivity, respectively. First, on this view we should ascribe to language models the ability to reason about beliefs (albeit to a lesser extent than humans); as others have noted (Kosinski, 2023), this would elevate their importance as social and intelligent agents in their own right. Second, if language experience alone is sufficient to develop the ability to reason about beliefs, this undermines claims that any innate endowment or social experience is necessary.

The alternative interpretation, which we call the **axiomatic rejection** position, holds that we should deny a priori that language models can display certain abilities—such as the ability to reason about the mental states of others—due to the nature of their constitution, for example, their lack of embodiment, grounding, agency, or embeddedness in an interactive social environment (Bender & Koller, 2020; Searle, 1980). If this view is correct, then GPT-3's success at the False Belief Task must be taken as evidence that the task itself is a flawed instrument for measuring the ability to attribute belief states (Raji, Bender, Paullada, Denton, & Hanna, 2021). On this view, LLMs' better-than-chance performance could perhaps be achieved by other means—e.g., an unintended Clever Hans effect, in which the LLM exploits unidentified confounds in the stimuli (Niven & Kao, 2019)—that do not reflect a capacity equivalent to the one the task was designed to measure. Accordingly, "passing" the test would constitute a kind of *reductio ad absurdum* of the test's validity; similar "proofs by absurdity" have been used to demonstrate potential flaws in other instruments, such as fMRI (Bennett, Miller, & Wolford, 2009) and measures of survey validity (Maul, 2017). Under this interpretation, the current results could be used to support existing critiques of the use of the False Belief Task (Bloom & German, 2000). An important secondary implication of the axiomatic rejection view is that no empirical behavioral evidence could resolve a debate about whether a given class of intelligent agents have the ability to reason about the mental states of others.

Of course, a third possibility would be to adopt a view that navigates between the duck test and axiomatic rejection positions. For example, one could argue that LLMs are indeed

a priori incapable of attributing and representing belief states (as in the axiomatic rejection view), but that this does not necessarily invalidate the utility of the tests for *human* subjects. This **differential construct validity** view is roughly the one adopted by Ullman (2023) in a response to related contemporary work (Kosinski, 2023). Considering this dilemma with respect to the broader question of theory of mind (ToM), Ullman (2023, p. 9) writes,

> …one can in principle hold the view that LLMs do not have ToM, while still thinking that ToM tests are valid when it comes to people…scholars have pointed out decades ago that people likely attribute intelligence not just on the basis of behavior but also on the basis of the algorithms and processes that generated that behavior.

This emphasis on internal states (as opposed to just behavior) is sometimes called *psychologism* (Block, 1981) and is typically seen as at odds with purely behaviorist or functionalist accounts of the mind (Block, 1980). If one adopts this *internalist* account of belief sensitivity, the question is thus whether LLMs and humans do indeed use different processes and mental representations to solve the False Belief Task. A further, deeper question is at what degree of granularity this issue of equivalent mental processes ought to be defined and operationalized. As Block (1980) notes, operationalization at the level of observable behavior (or high-level function) may be overly *liberal* in terms of which entities are granted mind-likeness—yet the functionalist rejoinder is that excessive specificity may be *chauvinistic* in terms of which entities it excludes.

While it may sound unlikely that LLMs use similar representations to solve the False Belief Task as humans, this is ultimately an empirical question and should be tested with further experimentation and probing. If this future work indicates that LLMs *do* in fact use similar processes as humans, researchers must then decide whether these processes ought to be described as belief attribution—in both humans and LLMs—or whether they are more appropriately characterized as emerging from a suite of domain-general, "lower-level" processes,for example, what Heyes (2014) calls *submentalizing*. Alternatively, if empirical probing uncovers distinct strategies in humans and LLMs, consistent with the **differential construct validity** view, researchers would be able to preserve both the False Belief Task as an instrument and the view that LLMs do not reason about belief states in a manner analogous to humans.

We do not explicitly endorse any of the competing interpretations presented here. In our view, there are persuasive arguments from all sides, and the evidence presented here and in related work (Kosinski, 2023) cannot adjudicate between them. As noted above, resolving this debate will require both greater refinement of the underlying *theoretical construct* (i.e., belief attribution) and the *instruments* used to measure it. This process may be informed by insights from work in comparative cognition, which we turn to below.

4.1.2. Insights from comparative cognition

A similar debate can be found in research on whether infants and nonhuman animals have the ability to attribute and represent belief states. For example, in recent years, evidence has accrued that certain great apes (e.g., chimpanzees) exhibit behavior *consistent with* this ability

(Hare, Call, Agnetta, & Tomasello, 2000; Krupenye & Call, 2019b; Krupenye, Kano, Hirata, Call, & Tomasello, 2016); however, there remains considerable debate over whether this evidence is necessarily indicative of the underlying capacity, or whether identical behavior could in principle be explained by other mechanisms (Halina, 2015; Penn & Povinelli, 2007). Certain aspects of this debate resemble the competing interpretations described above, namely the question of whether we ought to adopt an *intentional stance* with respect to non-human animals' ability to ascribe belief states (Dennett, 1987; analogous to the **duck test** position), or whether a more *deflationary* account involving domain-general, low-level processes (e.g., *submentalizing*) is more appropriate (Heyes, 2014).

In particular, one view emphasizes the fact that most evidence consistent with belief attribution (or "mindreading") in nonhuman animals is *also* consistent with the hypothesis that nonhuman animals are simply responding to observable behavioral regularities, without attributing or representing latent mental states at all. This view is sometimes called the "logical problem" (Halina, 2015; Lurz, 2009) and holds that this simpler null hypothesis—i.e., that nonhuman animals are engaging in "complementary behavior reading" rather than "mindreading"—must first be rejected (Penn & Povinelli, 2007; Povinelli & Vonk, 2004).

Of course, as Halina (2015) notes, any individual capable of belief attribution presumably still does so on the basis of observable behavior, which makes adjudicating between these competing interpretations (i.e., identifying veridical belief attribution) very challenging. Halina (2015, p. 485) suggests that the challenge can be surmounted by employing a range of different experiments with diverse techniques and distinct "observables":

> Doing so provides evidence for mindreading insofar as it establishes that subjects are responding to a diverse set of observable variables (eyes closed, opaque barrier present, head turned) as belonging to the same abstract equivalence class (situations that lead to a state of not seeing).

The fact that this issue is still under debate in the comparative cognition literature (Povinelli, 2020) suggests that resolving the question for LLMs will likely prove a serious challenge in the years to come; as noted in the previous section, it is also possible that the relevant philosophical theories (e.g., functionalism, psychologism, etc.) will prove impossible to adjudicate between Block (1980).

Moving forward, however, we argue that the strategy presented by Halina (2015) above seems like a promising and tractable approach: if LLMs exhibit behavior consistent with belief attribution in a wide range of experiments using a diverse class of stimuli, then it makes sense to ascribe to them the capacity to attribute and represent belief states—assuming, that is, that we would do the same for human participants in those same experiments (Dennett, 1978, 1987).

4.2. What can belief sensitivity tell us about theory of mind?

In the current work, we have restricted our claims to the question of sensitivity to true and false beliefs. However, this ability to represent the belief states of others is sometimes viewed

as part of a broader set of abilities—alternatively called mindreading, mentalizing, or theory of mind (Apperly, 2012). A question of growing concern in the field is whether theory of mind writ large is a coherent theoretical construct that offers explanatory value, or whether it is a convenient abstraction consisting of disparate, loosely related skills. One way to tackle this question is to consider the convergent and predictive validity of distinct instruments designed to measure theory of mind.

Theory of mind is an extraordinarily broad construct, and, accordingly, instruments have been designed to assess distinct components: reasoning about false beliefs (Wimmer & Perner, 1983), explaining or interpreting behavior in stories (Dodell-Feder et al., 2013; Happé, 1994), inferring emotional states from pictures (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), attributing mental states to animated shapes (Abell, Happe, & Frith, 2000), and more. Unfortunately, these tasks often display poor *convergent validity*: that is, performance on one task does not reliably correlate with performance on another (Gernsbacher & Yergeau, 2019; Gough, 2021; Hayward & Homer, 2017). Of course, tasks do converge in some cases; for example, performance on the short story task (Dodell-Feder et al., 2013) has been shown to correlate with performance on reading the mind in the eyes task (Baron-Cohen et al., 2001) (see Giordano et al., 2019; Dodell-Feder et al., 2013). However, the fact that convergent validity is so low in general (Gernsbacher & Yergeau, 2019; Hayward & Homer, 2017) suggests that these tasks are not, in fact, measuring the same thing—calling the coherence of "theory of mind" into question. As Gernsbacher and Yergeau (2019) note, some of these instruments also display poor predictive validity: that is, they do not reliably predict measures of behavior in social settings. Both facts are discouraging with respect to the question of whether theory of mind is a coherent construct: if tasks designed to measure it do not correlate with each other or with social behavior more generally, there is little justification for unifying these disparate abilities under a single "umbrella term."

While the empirical evidence presented in this paper clearly cannot settle this question, future work in this vein could contribute to the debate. Specifically, the performance of an LLM could be assessed across a *battery* of tasks designed to assess theory of mind (see also Kosinski, 2023), using a human benchmark in each case. Using this method, researchers could ask to what extent performance on each measure could be explained *in principle* by distributional statistics alone (as in the current work) and, crucially, to what extent these tasks display convergent validity in humans and LLMs. This would provide another robust test of the coherence of theory of mind as a construct; the results may help inform debates about whether we ought adopt a *pluralist* or *eliminativist* view of theory of mind (Gough, 2021), particularly when it comes to LLMs.

4.3. Using LLMs to study human comprehenders

The current work used GPT-3, an LLM, as a *baseline* for quantifying the extent to which human-level performance on the False Belief Task could be attributed to exposure to language statistics alone. This approach echoes past work using language models as "psycholinguistic subjects" (Futrell, Wilcox, Morita, & Levy, 2018; Jones et al., 2022; Linzen & Baroni, 2021; Michaelov, Coulson, & Bergen, 2022; Trott & Bergen, 2021) to investigate whether

distributional language statistics are sufficient *in principle* to explain human-level behavior at a task. Other contemporaneous work (Kosinski, 2023; Sap, LeBras, Fried, & Choi, 2022; Ullman, 2023) has asked whether LLMs exhibit evidence of theory of mind specifically. This LLM comparison approach allows us to empirically test theories that other sorts of innate capacities and learned experiences are *necessary* to display behavior consistent with belief attribution. However, there are important objections to using LLM performance to evaluate the claim that distributional information underlies human belief sensitivity *in practice*.

First, there are many differences between human comprehenders and LLMs which mean that the latter may not provide a psychologically plausible mechanism for the former. Some of these differences—the fact that humans are exposed to language in a rich social and multimodal context—might allow children to learn more from the same distributional information than models can. Our approach is designed to measure the sufficiency of distributional information in the absence of this scaffolding. Other differences, however, may artificially inflate estimates of how much could be learned by humans from language alone. Most notably, modern LLMs are trained on orders of magnitude more words than a human will see in their lifetime (Ullman, 2023). Children are estimated to be exposed to around 3–11 million words per year, for a total of 30–110 million words by the time they reach adult-like linguistic competence at age 10 (Hart & Risley, 1992; Hosseini et al., 2022). By contrast, GPT-3—the model used in our analysis—has been exposed to more than 200 billion words: $\sim$ 2000 times that of a 10 year old (Warstadt & Bowman, 2022).

If this scale of data is necessary to learn the nuanced statistical contingencies required for successful performance at a task, it would undermine the inference that LLM performance is indicative of what humans could do with distributional information. Importantly, however, LLM performance is also related to the number of parameters in the model. In our exploratory analyses, we found that the largest GPT-3 models tested (*text-davinci-002*) performed much better on the False Belief Task than smaller models, consistent with past work suggesting that LLMs may obey certain "scaling laws" (Kaplan et al., 2020). The differential effects of model parameter count and training dataset size on False Belief Task performance remain unclear; very large models (or a human brain with hundreds of trillions of synapses) could potentially perform similarly to GPT-3 using less data.[2] A useful approach here would be to compare the predictive power of LLMs trained with different amounts (and sources) of training data to determine whether a "developmentally realistic" amount of training data could yield behavior consistent with the capacity in question (Hosseini et al., 2022).

These questions will become even more critical in the near future. LLMs will continue to increase in size and will be trained on larger datasets—orders of magnitude more words than a human is exposed to in a lifetime.[3] If machines behave indistinguishably from humans on these tasks, the question of whether achievement on the False Belief Task itself constitutes sufficient evidence for false belief sensitivity will raise deep philosophical questions for the field: should such LLMs be considered "agents" capable of reasoning about the belief states of others, or should these demonstrations force us to reevaluate the utility of the instruments we use to measure these cognitive capacities?

Even if developmentally plausible models do show humanlike behavior at a task, this does not imply that humans are using the same statistical mechanism as these models. There could be multiple distinct routes to the same behavior, and humans could in fact be using innate or domain-specific mentalizing capacities to produce behavior that models learn to imitate from language statistics. Even insofar as humans do use distributional information to make inferences about beliefs, there are a variety of plausible mechanisms for this, including domain-general statistical learning (Aslin, 2017), language-specific predictive processing (Heilbron, Armeni, Schoffelen, Hagoort, & De Lange, 2022), and innate but non-statistical inferential mechanisms (Penn, Holyoak, & Povinelli, 2008). The specific mechanistic theory operationalized by LLMs is that humans use language statistics to predict upcoming input. Results showing that LLM representations can predict up to 100% of explainable variance in brain activity have been taken as evidence for this hypothesis (Schrimpf et al., 2021). However, Antonello and Huth (2022) show that statistical language representations learned for other objectives (e.g., translation) are similarly predictive of human brain responses, implying that the correlation of human and LLM data may be due to features of language statistics generally rather than a close mechanistic similarity. In order to adjudicate between these accounts, researchers will need to identify and empirically test divergent predictions of these mechanistic accounts.

One benefit to using LLMs as an *operationalization* of a theory is that, as models, they offer more opportunities for testing various more specific mechanisms or hypotheses. For example, what kinds of language input are most critical for developing the ability to reason about mental states? Past work has argued for the importance of at least three distinct sources, including exposure to mental state verbs (Brown et al., 1996), the structure of interactive conversation (Harris, 2005), and certain syntactic constructions (Hale & Tager-Flusberg, 2003). Future work could compare different models with different training corpora (e.g., primarily dialogue vs. essays) to help isolate how much information is provided by each source of linguistic experience.

While these objections highlight the importance of future theoretical and empirical work, we believe that evidence for the *sufficiency* of distributional information for competent False Belief task performance is a critical step toward assessing the plausibility of experience-based theories of belief attribution in humans.

5. Conclusion

Where does the human ability to reason about beliefs of others come from? It could emerge in part from an innate, biologically evolved capacity (Bedny et al., 2009). It might also depend on experience, including language input (de Villiers & de Villiers, 2014). The current results help quantify the contribution of language input. On a text-based version of the False Belief Task, humans responded correctly (i.e., in a manner congruent with a character's belief states) 82.7% of the time, while the largest LLM tested responded correctly 74.5% of the time; additionally, LLM behavior did not fully explain human behavior. This suggests that language statistics alone are sufficient to generate *some* sensitivity to false belief but, crucially, not to

fully account for *human* sensitivity to false belief. Thus, the ability of humans to attribute mental states to others may involve *linking* this linguistic input to innate capacities or to other embodied or social experiences.

Open Research Badges


This article has earned Open Data, Open Materials badges, and pre-registered. Data and materials are available at https://osf.io/hu865/ and pre-registered are available at https://osf.io/zp6q8, https://osf.io/agqwv.

Notes

1 Recently, pre-trained language models have been fine-tuned using Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). This process arguably leads to a training signal that is not purely based on distributional language statistics and so we do not use RLHF models in this analysis.
2 While it is difficult and problematic to compare the computational power of neural networks and human brains, an estimated 1.5×10^{14} synapses in the human adult neocortex (Drachman, 2005) is $\sim$ 850 times the number of parameters in GPT-3 (1.75×10^{11})
3 Indeed, in the course of revising this article, GPT-4 was released, achieving substantially higher scores on a range of different psychometric tests (OpenAI, 2023a).

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 109–132). Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/2021.conll-1.9

Abell, F., Happe, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16.

Antonello, R., & Huth, A. (2022). Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*. Advance online publication. https://doi.org/10.1162/nol_a_00087

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825–839.

Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1–2), e1373.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

Bedny, M., Pascual-Leone, A., & Saxe, R. R. (2009). Growing up blind does not change the neural bases of theory of mind. *Proceedings of the National Academy of Sciences*, *106*(27), 11312–11317.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5185–5198). Stroudsburg, PA: Association for Computational Linguistics.

Bennett, C. M., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for multiple comparisons correction. *Neuroimage*, *47*(Suppl 1), S125.

Block, N. (1980). Troubles with functionalism. In *The language and thought series* (pp. 268–306). Harvard University Press.

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, *90*(1), 5–43.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31.

Bradford, E. E., Brunsdon, V. E., & Ferguson, H. J. (2020). The neural basis of belief-attribution across the lifespan: False-belief reasoning and the n400 effect. *Cortex*, *126*, 265–280.

Brown, J. R., Donelan-McCall, N., & Dunn, J. (1996). Why talk about mental states? the significance of children's conversations with friends, siblings, and mothers. *Child Development*, *67*(3), 836–849.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

de Villiers, J. G., & de Villiers, P. A. (2014). The role of language in theory of mind development. *Topics in Language Disorders*, *34*(4), 313–328.

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, *1*(4), 568–570.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: A new task for assessing theory of mind in adults. *PLoS ONE*, *8*(11), e81279.

Drachman, D. A. (2005). Do we have brain to spare? *Neurology*, *64*(12), 2004–2005.

Fairchild, S., & Papafragou, A. (2021). The role of executive function and theory of mind in pragmatic computations. *Cognitive Science*, *45*(2), e12938.

Futrell, R., Wilcox, E., Morita, T., & Levy, R. (2018). RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. arXiv preprint. arXiv:1809.01329

Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology*, *7*(1), 102.

Giordano, M., Licea-Haquet, G., Navarrete, E., Valles-Capetillo, E., Lizcano-Cortés, F., Carrillo-Peña, A., & Zamora-Ursulo, A. (2019). Comparison between the short story task and the reading the mind in the eyes test for evaluating theory of mind: A replication report. *Cogent Psychology*, *6*(1), 1634326.

Gough, J. (2021). Does the neurotypical human have a 'theory of mind'? *Journal of Autism and Developmental Disorders*, *53*(2), 853–857.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, *6*(3), 346–359.

Halina, M. (2015). There is no special problem of mindreading in nonhuman animals. *Philosophy of Science*, *82*(3), 473–490.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771–785.

Harris, P. L. (2005). Conversation, pretense, and theory of mind. In *Why language matters for theory of mind* (pp. 70–83). New York, NY: Oxford University Press.

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, *28*(6), 1096.

Hayward, E. O., & Homer, B. D. (2017). Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology*, *35*(3), 454–462.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131–143.

Hosseini, E. A., Schrimpf, M. A., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. bioRxiv, 2022–10.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development*, *76*(2), 356–370.

Johnson, S. (2022). A.I. is mastering language. Should we trust what it says? *The New York Times*. https://www.nytimes.com/2022/04/15/magazine/ai-language.html

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. (2022). Distributional semantics still can't account for affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *44*, pp. 482–489). Austin, TX: Cognitive Science Society.

Jurafsky, D., & Martin, J. H. (2014). Speech and language processing. Vol. 3. https://web.stanford.edu/~jurafsky/slp3/

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint. arXiv:2001.08361.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint. arXiv:2302.02083.

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

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

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

Leslie, A. M. (2001). Theory of mind. In N. J. Smelser& P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 15652–15656). Oxford, England: Pergamon

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1813–1827). Stroudsburg, PA: Association for Computational Linguistics.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Lurz, R. (2009). If chimpanzees are mindreaders, could behavioral science tell? Toward a solution of the logical problem. *Philosophical Psychology*, *22*(3), 305–328.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*. Advance online publication. https://doi.org/10.1109/TCDS.2022.3176783

Niven, T., & Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. arXiv preprint. arXiv:1907.07355.

OpenAI (2023a). GPT-4 technical report. arXiv preprint. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

OpenAI (2023b). OpenAI model documentation. Retrieved from https://platform.openai.com/docs/models/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint. arXiv:2203.02155.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109–130.

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

Pluta, A., Krysztofiak, M., Zgoda, M., Wysocka, J., Golec, K., Wójcik, J., Włodarczyk, E., & Haman, M. (2021). False belief understanding in deaf children with cochlear implants. *Journal of Deaf Studies and Deaf Education*, *26*(4), 511–521.

Povinelli, D. J. (2020). Can comparative psychology crack its toughest nut. *Animal Behavior and Cognition*, *7*(4), 589–652.

Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language*, *19*(1), 1–28.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). Ai and the everything in the whole wide world benchmark. arXiv preprint. arXiv:2111.15366.

Sap, M., LeBras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? On the limits of social intelligence in large LMs. arXiv preprint. arXiv:2210.13312.

Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage*, *101*, 268–275.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In *New essays on belief* (pp. 75–99). Berlin, Germany: Springer.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Shevlin, H. (2022). Uncanny believers: Uncanny believers: chatbots, beliefs, and folk psychology. Unpublished manuscript. https://henryshevlin.com/wp-content/uploads/2021/11/Uncanny-Believers.pdf

Sinclair, A., Jumelet, J., Zuidema, W., & Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, *10*, 1031–1050.

Trott, S., & Bergen, B. (2021). Raw-C: Relatedness of ambiguous words–in context (a new lexical resource for English). arXiv preprint. arXiv:2105.13266.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. arXiv preprint. arXiv:2302.08399.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008). Red Hook, NY: Curran Associates Inc.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In S. Lappin& J,-P. Bernardy (Eds.), *Algebraic structures in natural language* (pp. 17–60). Boca Raton, FL: CRC Press

Webb, M. A., & Tangney, J. P. (2022). Too good to be true: Bots and bad data from Mechanical Turk. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916221120027

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

Xie, J., Cheung, H., Shen, M., & Wang, R. (2018). Mental rotation in false belief understanding. *Cognitive Science*, *42*(4), 1179–1206.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.