PSYCHOLOGICAL SCIENCE

Quantitative mental state attributions in language understanding

Julian Jara-Ettinger[1,2]* and Paula Rubio-Fernandez[3,4]

Human social intelligence relies on our ability to infer other people's mental states such as their beliefs, desires, and intentions. While people are proficient at mental state inference from physical action, it is unknown whether people can make inferences of comparable granularity from simple linguistic events. Here, we show that people can make quantitative mental state attributions from simple referential expressions, replicating the fine-grained inferential structure characteristic of nonlinguistic theory of mind. Moreover, people quantitatively adjust these inferences after brief exposures to speaker-specific speech patterns. These judgments matched the predictions made by our computational model of theory of mind in language, but could not be explained by a simpler qualitative model that attributes mental states deductively. Our findings show how the connection between language and theory of mind runs deep, with their interaction showing in one of the most fundamental forms of human communication: reference.

INTRODUCTION

People's behavior is rich with information about their mental life. A subtle yawn can betray that your friend is bored or tired; a glance at their wristwatch might suggest that they are eager to leave; or a pause before answering a sensitive question can reveal that they are considering how to reply. Inferences like these are fundamental to our everyday lives, allowing us to understand other people's behavior, determine what to expect, and decide how to react. To what extent can we make these inferences in a precise and fine-grained manner based on how people speak?

People's ability to infer each other's mental states—known as a theory of mind—has been historically studied in the context of physical action. By attending to how agents move and behave, people can infer a range of mental states including goals, preferences, and knowledge (*1*–*9*). However, much of real-world social behavior happens in the context of linguistic interactions, where people's words can reveal the contents of their minds, even in the absence of physical cues (e.g., when speaking on the phone).

In these conversational contexts, speakers often willingly disclose their mental states by using mentalistic words (*10*), such as when we confirm that we understand something or confess that we are confused and feel embarrassed. However, even the most basic nonmentalistic words, such as articles and adjectives, can reveal aspects of what a speaker wants or knows. For instance, if a friend asked you to bring "the blue cup" from the kitchen, their words might suggest that they expect you to find only one such cup among several others or that you know which cup they are talking about.

Despite the mental state information available in speakers' nonmentalistic words, listener mental state reasoning in simple communicative tasks appears to be unexpectedly limited (*11*–*15*) [cf. (*16*–*18*)]. For instance, when a speaker who cannot see the smallest of three balls requests "the small ball," listeners do not immediately take the middle-sized ball (the smallest one from the speaker's perspective). Instead, people often first look at—and sometimes even reach for—the ball that the speaker is unaware of.

Critically, these mental state reasoning failures emerge when listeners are explicitly told about the speaker's perspective and must interpret what the speaker says accordingly. In more realistic interactions, however, we are rarely told how our interlocutor's perspective differs from our own so that we can interpret their words accordingly. Instead, we often do the reverse: We infer what our interlocutor knows (or does not know) during the course of our conversations, based on what they say and how they say it.

Here, we sought to test people's ability to extract speakers' knowledge from their choice of words (rather than using speakers' knowledge to interpret their words). When attributing mental states from observable action, people's inferences are nuanced and quantitative (*1*, *3*, *4*, *6*), similar to those characteristic of low-level processes such as perception and motor control [e.g., consider the precision needed to move one's arm and swiftly pick up a hat (*19*–*21*)]. However, it is unknown whether this level of granularity might extend to mental state inferences based on simple word choices. While past work has found that people can attend to mental state information in communicative interactions (*16*–*18*), these results do not reveal whether these inferences are coarse and qualitative, or nuanced and fine-grained. Our goal was therefore to test whether listeners can derive sophisticated mental state inferences from speakers' minimal linguistic choices, which they can then deploy as needed (such as to understand the speaker's message or determine whether they are aware of a particular piece of information; see discussion).

To test this, we used a more complex, yet perhaps more natural task than those typically used to probe mental state reasoning in language comprehension: Participants had to infer both the speaker's referential intent and their knowledge from their choice of words. This paradigm better reflects the structure of normal communicative interactions, where speakers' knowledge and referential intent are both unobservable and must be inferred in tandem.

We focused our study on one of the simplest and most central linguistic events where language and mental state reasoning interface: reference production and resolution. In deciding what to call an object, speakers must be aware of what listeners will treat as a potential referent. For instance, if there was a single cup in the

[1]Department of Psychology, Yale University, New Haven, CT, USA. [2]Department of Computer Science, Yale University, New Haven, CT, USA. [3]Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo, Norway. [4]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA.
*Corresponding author. Email: julian.jara-ettinger@yale.edu

kitchen, your friend may ask you to bring them "the cup." However, if multiple cups are in view, they should recognize that the bare description is ambiguous and add additional information such as the cup's color or size. Conversely, listeners can infer what speakers may or may not know based on their choice of words. If, for instance, your friend asked you for "the cup" when two cups are in view, you could infer that they have one of the cups in mind but did not realize there was a second one (otherwise they would have asked you for "a cup"). Likewise, if your friend produced a modified description, such as "the big cup," you could infer that they are aware that there is more than one cup in view and that not all cups are the same size. Given these potential inferences, we used referential communication to investigate whether listeners can derive fine-grained mental state inferences from speakers' use of nonmentalistic words.

To evaluate listeners' capacity to derive theory of mind inferences from speakers' choice of words, we developed a computational model that derives precise and nuanced mental state inferences in a quantitative manner. These inferences in our model not only identify what an agent may or may not know but also provide fine-grained levels of confidence over these inferences. Our computational model therefore allows us not only to test whether people can infer mental states from speakers' choice of words but to also see whether these inferences are quantitative. Alternatively, mental state inferences from speakers' word choices might be coarse and qualitative, similar to those arising from heuristics and biases that are broadly correct but lack nuance [e.g., when complex decisions are influenced by the problem's framing or by anchoring effects (*22*, *23*)]. To explore this latter possibility, we contrasted our model with a simpler deductive model that determines reference and knowledge deductively based on speakers' literal descriptions, without reasoning about their choice of words.

In experiment 1, we first tested people's ability to infer speakers' mental states based on how they use color adjectives and whether these inferences are best explained by our quantitative theory of mind model or by our alternative deductive model. In linguistic events, however, mental state inferences must be adjusted to speakers' individual communicative patterns (*24*–*26*). In experiment 2, we thus tested if, like our model, people adjust their inferences based on evidence of speakers' propensity to use adjectives redundantly (i.e., when they are not necessary to preempt an ambiguity). Last, in experiment 3, we tested more complex visual displays (using pictures of real-world objects) where speakers dynamically use different adjective types and more or less specific words (manipulating color modification, size modification, and noun choice within participants).

Computational framework

While substantial computational work has looked at how people identify speakers' referential intent (*27*–*30*), including cases where speakers' knowledge affects how they speak (*31*, *32*), this work has primarily focused on situations where the ultimate goal is to resolve reference. Here, rather than focusing on how knowledge of mental states affects language understanding, we focus on how language understanding supports inferences about mental states.

We take as a starting point advances in computational cognitive science showing how human social reasoning can be understood as Bayesian inference over a mental model of a rational agent in linguistic (*27*, *28*), pedagogical (*33*–*35*), and nonlinguistic interactions (*1*, *4*, *6*). Under these frameworks, observers infer an actor's mental states by considering what types of beliefs and desires would lead a rational agent to act, speak, or communicate in the observed manner. Our model falls within this framework: Given an utterance, we perform a joint inference over the mental states and referential intent that combined explain speakers' choice of words.

To illustrate the logic of our model, consider a situation like the one in Fig. 1A. Here, a speaker describes one of the four shapes in each of the two displays ("the square" in the left-side display and "the green triangle" in the right-side display). However, the speaker has a blind spot and cannot see one of the four cells, incorrectly believing that each display only has three shapes to choose from. Given the referential expressions in Fig. 1A, we can infer that the speaker's blind spot must be one of the top cells. Listeners, however, can often go even further than this logical deduction. In Fig. 1B, for instance, the speaker used color adjectives in both displays, helping the listener identify the intended shape and discard the alternative shape in the top-left cell. This suggests that the speaker could see that cell (despite never directly referring to it) and that their blind spot must therefore be the top-right cell. In other cases, however, such as in Fig. 1F, the speaker may be using color words redundantly, and our model also aims to capture how listeners must adjust their inferences accordingly.

We formalize the logic of this inference by building on past work showing that both reference resolution and mental state inferences are instantiated as Bayesian inference over models of a rational speaker (*27*, *28*) or a rational actor (*1*, *4*, *6*), respectively. Within a probabilistic framework, we can express the problem of jointly inferring speakers' beliefs and intended referent as computing the posterior distribution (see the Supplementary Materials for a detailed derivation of Eq. 1)

$$p(t,b \mid u) \propto \sum_{r\in[0,1]} p(u \mid t,b,r)\, p(r)\, p(t \mid b)\, p(b) \tag{1}$$

Here, t is the speaker's intended referent (or target), formalized as one of the objects in the visual scene; b is the speaker's belief and represents the objects that the speaker is aware of, formalized as any subset (including the full set) or objects in the scene; r is the speaker's unknown propensity to speak redundantly; and u is the speaker's utterance. In our tasks, we used a uniform distribution over beliefs where one of the objects is hidden from the speaker [i.e., $p(b) = 1/4$ for each belief where three objects are known] and a uniform distribution over $p(t \mid b)$ [i.e., $p(t \mid b) = 1/3$ for each observable object].

In line with models of mental state attribution and reference resolution (*3*, *6*, *27*, *36*), the probability that the speaker produces utterance u [$p(u \mid t, b, r)$] is obtained by assuming that the speaker is motivated to be as informative as possible (*37*) adjusted with the empirical finding that speakers often overspecify (*38*–*41*).

To calculate this likelihood term, we first assume that the speaker has a fixed and known probability of accidentally producing an underinformative expression (e.g., "the triangle" when two triangles are in view), estimated in a separate task ($P = 0.055$ and 0.056 for experiments 1 and 2, respectively, and $P = 0.047$, 0.163, and 0.217 for color, size, and category in experiment 3; see the Supplementary Materials for experiment details). In the remaining cases, the speaker selects the shortest sufficiently informative utterance with probability $1 - r$ and introduces a redundant adjective with probability r. Critically, however, we treat this probability as variable across speakers. Thus, rather than using a single parameter, we represented the

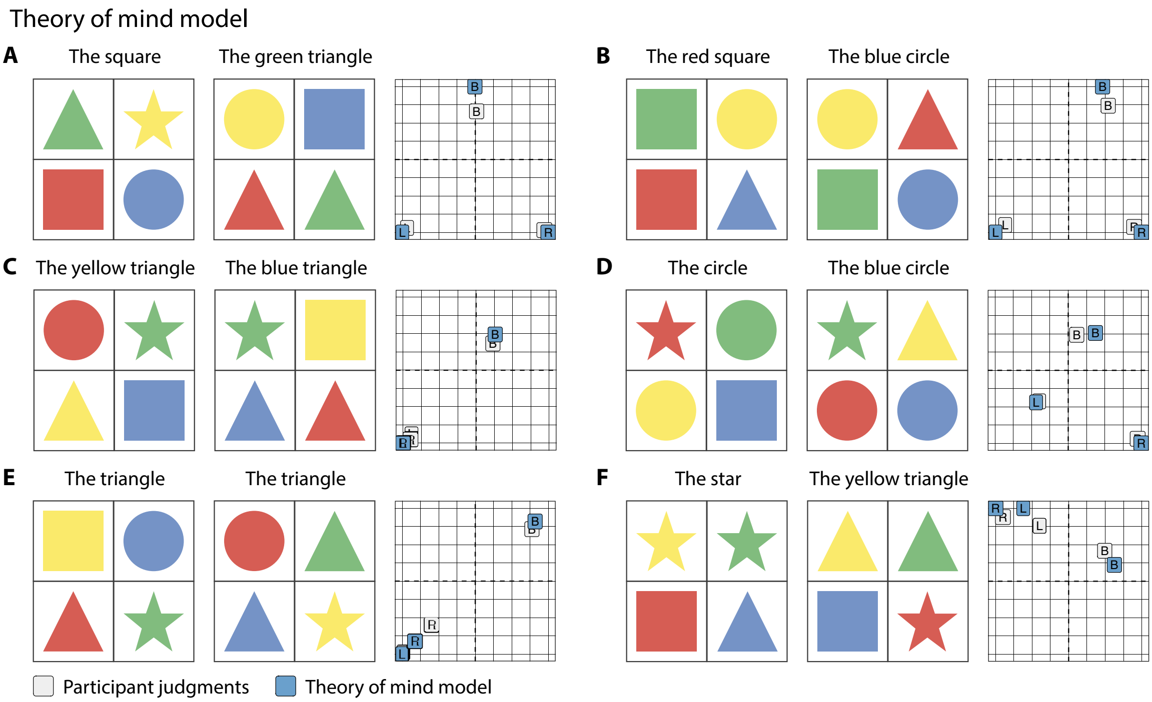

Fig. 1. Example trials from experiment 1 along with participant judgments and our theory of mind model's predictions. (**A** to **F**) Experiment trials, consisting of two visual scenes, each with a speaker utterance (presented on top). For each panel, trackpads to the right of each panel show average participant judgments and model predictions. L marks the inferred referent on the left-side display, R the inferred referent on the right-side display, and B the inferred blind spot.

probability of overspecification as a Beta distribution with parameters estimated in a separate task [using $B(0.39,0.32)$ and $B(0.32,1.12)$ for experiments 1 and 2, respectively, and $B(0.43,0.82)$, $B(0.28,1.18)$, and $B(0.29,0.22)$ for color, size, and category in experiment 3; see the Supplementary Materials for experiment details]. Thus, the likelihood function captures the idea that speakers tend to produce the shortest possible expression, with a small fixed probability of underspecifying and an unknown speaker-variable probability of overspecifying.

Our theory of mind model generates quantitative predictions that reflect listener beliefs about the speaker's intended referent and knowledge, assigning probabilities to each potential referent and to each potential belief. To contrast these predictions with a more qualitative account—i.e., an account where listeners infer agents' knowledge, but lack quantitative estimates of their certainty—we also considered a simpler model that infers intended referents and knowledge without considering speakers' choice of words. Because, in contexts where belief inference is not at stake, reference resolution is well captured as probabilistic inference over speaker intentions (*27*, *28*), our alternative model preserved our main model's probabilistic framework, with the difference that it ceased to consider how knowledge influences a speakers' choice to include or omit adjectives. Formally, we achieved this by simply placing a uniform distribution over utterances that describe the target in $p(u \mid t, b, r)$. Thus, this model captures a form of qualitative theory of mind: It understands that speakers will only describe objects that they know about [encoded in $p(t \mid b)$; Eq. 1] and uses this knowledge to deductively identify the speaker's blind spot, but it does not treat the use or omission of adjectives as carrying information about the speaker's mental states.

This deductive model makes identical predictions to our theory of mind model in some situations, but diverges in critical ways in others. In Fig. 2 (A and D), for instance, the deductive model can also identify the two referents correctly, but only in Fig. 2A does the deductive process reveal the probable location of the blind spots. In Fig. 2 (B and C), however, the deductive model is unable to infer the referents or the speaker's blind spot, as a full joint inference is required to understand what the speaker intends to say and what they know.

RESULTS

In experiment 1, we first tested people's ability to perform joint inferences about speakers' intended referents and knowledge, and whether these judgments matched the quantitative resolution captured in our theory of mind model. In experiment 2, we further tested whether participant inferences were sensitive to speaker-specific communicative styles, providing evidence that people adjust their mental state inferences not only to the content of the utterance but also to the speaker's propensity to use descriptive language. Last, in experiment 3, we extended our findings to test whether these mental state inferences remain when using complex real-world objects and speakers use more variable descriptions, including different nouns.

Experiment 1: Joint referent and belief inferences

In experiment 1, we used a language comprehension task based on a standard paradigm (*11*, *16*, *18*, *42*, *43*), extended to include multiple events where speakers' utterances revealed partial information about their mental states. For this first test, we used color modification

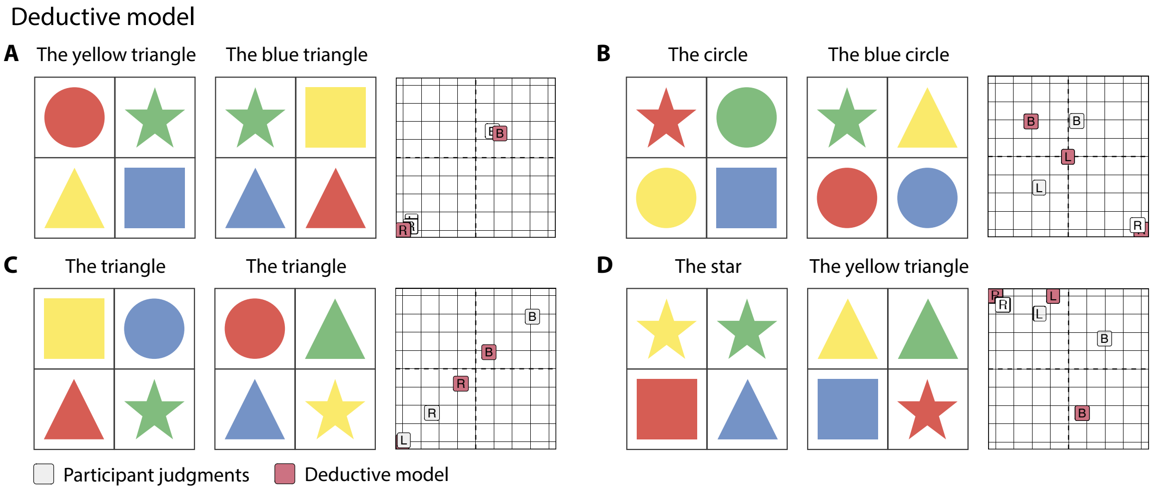

Fig. 2. Example trials from experiment 1 along with participant judgments and the deductive model's predictions. (**A**) to (**D**) show the same trials from Fig. 1 (C to F).

as a cue to speaker beliefs because color is often overspecified (*39*, *44*). Therefore, speakers' inclusion or omission of color adjectives cannot be treated as simple cues to knowledge, and listeners must consider the pattern of usage in the visual context to infer the speaker's knowledge. Participants were simultaneously presented with pairs of linguistic events like those in Fig. 1 (28 pairs of displays total; see the Supplementary Materials) and had to infer speakers' intended referent in each event and the blind spot shared across both events. Participants used three separate two-dimensional (2D) continuous trackpads to report their inferences and certainty (i.e., the closer they moved the marker toward a corner, the greater their certainty about the identified cell; see Fig. 1 and the Supplementary Materials for details).

Participant judgments showed a high fit to our quantitative theory of mind model, with a correlation of $r = 0.95$ for belief inferences [95% confidence interval ($CI_{95\%}$): 0.92 to 0.98] and a correlation of $r = 0.99$ ($CI_{95\%}$: 0.99 to 1.00; Fig. 3, A and B) for referent inferences. See Fig. 1 (A to F) for six example trials. The fact that people were not only able to infer speakers' mental states qualitatively but also to shift these judgments with the fine-grained precision characteristic of action understanding tasks suggests that participants were indeed able to perform quantitative theory of mind inferences from these simple linguistic events. To further evaluate this possibility, we considered our alternative model that performed deductive mental state inferences. This model's referent inferences also matched participant judgments with high precision ($r = 0.99$; $CI_{95\%}$: 0.99 to 1.00; Fig. 3, A and B) and were comparable to the inferences of our theory of mind model ($\Delta r = 0.001$; $CI_{95\%}$: −0.004 to 0.007). By contrast, the deductive model's belief inferences were markedly lower ($r = 0.56$; $CI_{95\%}$: 0.38 to 0.82) and outperformed by our theory of mind model ($\Delta r = 0.386$; $CI_{95\%}$: 0.15 to 0.56).

Figure 2 (B to D) (compare to Fig. 1, D to F, for our theory of mind's model predictions on the same displays) shows three example trials where participants readily combined information from both displays to jointly infer the speaker's intended referents and belief, while the deductive model produced inferences that failed to extract information available by thinking about the speaker's word choice (see the Supplementary Materials for trial-by-trial plots). In Fig. 2D, for instance, the deductive model performs similar to participants when identifying the referents. When inferring beliefs, however, participants infer that the speaker most likely cannot see the top-right cell based on their choice of words (in particular, the failure to use a disambiguating color word in the first event). By contrast, the deductive model infers that the blind spot is probably on the bottom cells because the speaker's referents always appear to identify objects in the top cells.

Similar to previous modeling work probing mental state inferences, our analyses focused on average judgments per trial (*1*–*4*, *6*, *9*). This allowed us to reveal the shared signal across participant judgments while removing potential noise introduced by requesting participants to report explicit inferences on a trackpad. Consistent with this, individual participant belief inferences showed an average $r = 0.59$ correlation with our theory of mind model, and a significantly lower $r = 0.35$ correlation with our deductive model [$t(59) = 5.43$, $P < 0.0001$ by paired t test]. Similar to our main results, a subject-level analysis showed no difference across models on referent inferences [$r = 0.93$ for both models; $t(59) = 0.38$, $P = 0.71$ by paired t test].

Experiment 2: Adjusting mental state inferences based on speaker's communicative style

The results from experiment 1 show that people can infer speakers' knowledge based on their adjective use. In that task, inferences relied only on general expectations of how often speakers use adjectives redundantly or contrastively (*38*, *39*, *45*, *46*). For these inferences to be accurate, however, they must be sensitive to different speakers' propensity to use adjectives redundantly. In experiment 2, we thus sought to test whether people adjust their mental state inferences by learning speaker-specific preferences on adjective use.

Experiment 2 was conceptually similar to experiment 1, where a hypothetical speaker with a blind spot described shapes in different visual displays. However, in contrast to experiment 1, experiment 2 now used five consecutive trials with the same speaker, allowing participants to learn speakers' propensity to use redundant adjectives. In addition, experiment 2 used size adjectives that, unlike color adjectives, have a default contrastive interpretation (*47*). Therefore, a

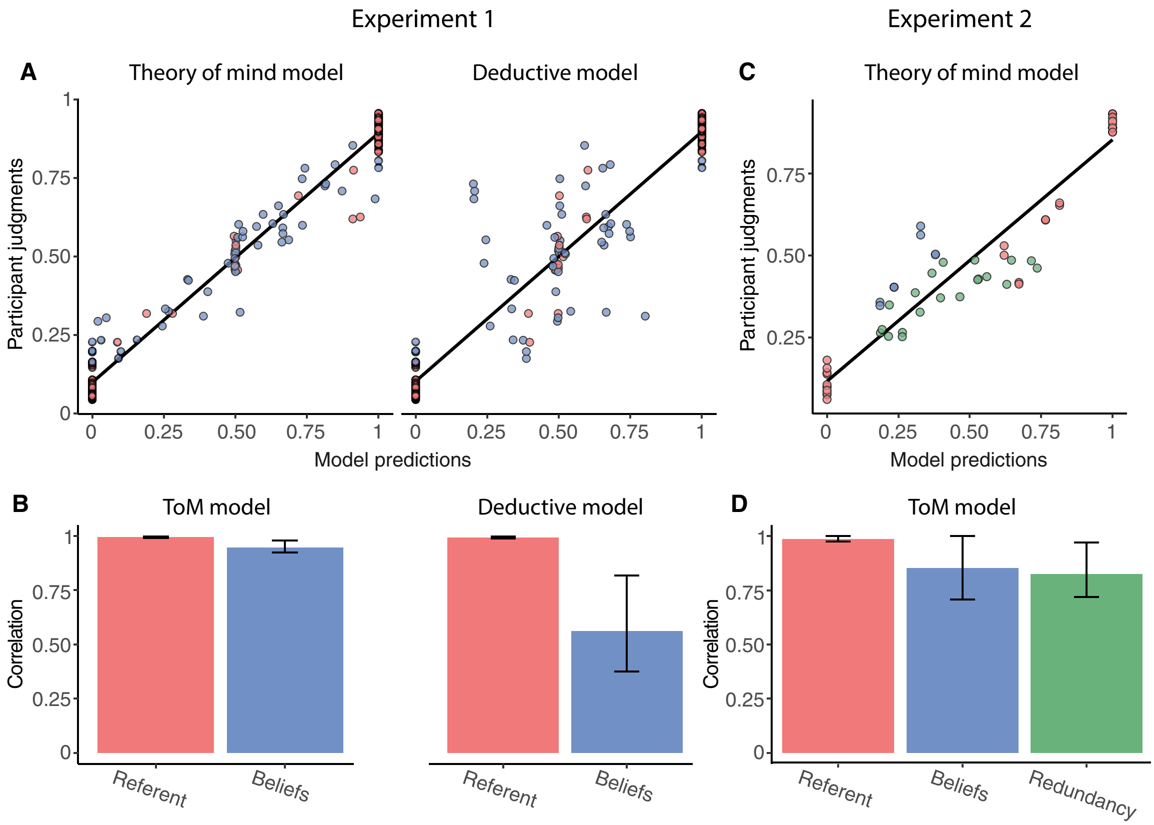

Fig. 3. Overall results for experiments 1 and 2. (**A**) Results from experiment 1. Each dot represents a judgment with model prediction on the *x* axis and average participant judgments on the *y* axis. Color indicates inference type (referent or belief). (**B**) Experiment 1 correlations as a function of model and inference type. (**C**) Results from experiment 2. (**D**) Experiment 2 correlations as a function of inference type. Black vertical bars show 95% bootstrapped CIs. ToM, theory of mind.

failure to treat a size adjective as revealing speakers' knowledge of the contrast object is a conservative test for how participants adjust their inferences to speaker-specific communicative styles.

All participants first completed three consecutive trials where the speaker alternated between referring to the top-right and the bottom-left cells, revealing that their blind spot must be the top-left or the bottom-right cell (actual cell positions randomized across participants; Fig. 4A). The descriptions of the targets were always unambiguous, but speakers varied in their propensity to use redundant size adjectives, ranging from being maximally succinct (0/3 redundant uses) to maximally redundant (3/3 redundant uses). In the fourth trial (last display in Fig. 4A), the speaker always used a size adjective that, if interpreted contrastively, revealed that they could see the bottom-right cell (although the likelihood of a contrastive use depended on the speaker's redundancy in the previous trials). In each of these four trials, participants were asked to identify the speaker's referent using a 2D trackpad and to rate the speaker's propensity to speak redundantly (using a continuous slide bar; see Materials and Methods).

In the fifth and critical trial, participants had to jointly infer the speaker's intended referent and their blind spot. Here, speakers always produced the ambiguous description "the star," which could refer to the top-left or bottom-right cells. If participants treated the adjective in the fourth trial as contrastive (inferring that the speaker sees the bottom-right cell), they should identify the bottom-right cell as the referent and the top-left one as the blind spot. As predicted, this pattern appeared when the speaker was maximally succinct (leftmost pair in Fig. 4, B and C). If, instead, participants treated the adjective in the fourth trial as a redundant use, then they should be unable to identify the referent or the blind spot in the fifth trial. We obtained this predicted pattern when the speaker was maximally redundant (rightmost pair in Fig. 4, B and C), with the intermediate conditions revealing a graded transition.

Consistent with the qualitative distributions, our quantitative theory of mind model showed a strong fit to participant judgments ($r = 0.96$; $CI_{95\%}$: 0.94 to 0.98; Fig. 3C), and this correlation was comparably strong in each component of referent inferences ($r = 0.99$; $CI_{95\%}$: 0.98 to 1.00), redundancy tracking ($r = 0.82$; $CI_{95\%}$: 0.72 to 0.97), and belief inferences ($r = 0.85$; $CI_{95\%}$: 0.71 to 1.00; Fig. 3D). Note that participants here made a single mental state inference, and we therefore cannot perform subject-level model correlations.

The pattern we observed in experiment 2 also provides conclusive evidence against our deductive model. By not being able to consider speakers' choice of words, the deductive model cannot learn speaker-specific levels of redundancy, cannot make any mental state inferences on the fourth trial, and can never derive the joint inference on the fifth trial (as a consequence, the model predictions have no variance, and it is not possible to compute a correlation between model predictions and participant judgments; see the Supplementary Materials for details).

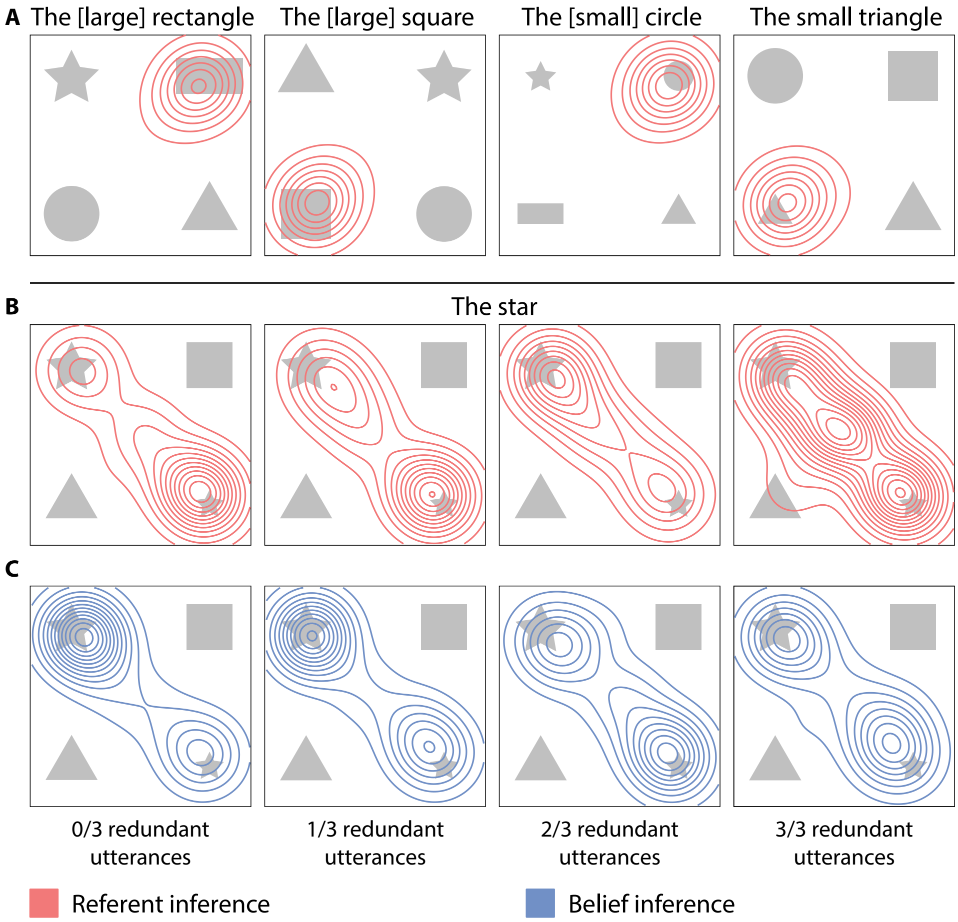

Fig. 4. Experiment 2 inference distributions. (**A**) Trials 1 to 4 in experiment 2. The text above each display shows the speaker's description. The use of size adjectives in brackets varied across conditions (from 0/3 to 3/3 redundant uses). The red rings show the distribution of selections on the trackpad. (**B** and **C**) Referent inferences (red) and knowledge inferences (blue) on the fifth trial as a function of speaker's redundancy. When the speaker was maximally succinct (leftmost pair), participants inferred that the ambiguous description referred to the bottom-right cell (making the top-left cell the blind spot); when the speaker was maximally redundant (rightmost pair), participants were unable to identify the referent or the blind spot.

Experiment 3: Joint referent and belief inferences from different adjective types and words

While the results from experiments 1 and 2 show that people can make nuanced mental state inferences from how people speak, they also leave three questions open. First, these studies used simple geometrical shapes as referents, enabling us to manipulate key contrasts while controlling for visual features. Would these results generalize to natural real-world objects? Second, experiments 1 and 2 focused on people's choice of color and size adjectives, respectively. However, people's knowledge is often also reflected in their noun choice rather than adjective use (e.g., a person aware of two dogs in a scene might choose to call the target dog "the Labrador" to avoid ambiguity). Last, the strength of mental state inferences depends on how often people use redundant adjectives. Can people flexibly adjust their expectations within a single task that varied adjective type? To answer these questions, we conducted a simplified replication of experiment 1 using pictures of real-world objects, where we varied color modification, size modification, and noun choice within participants.

Overall, participant inferences about speaker knowledge showed a correlation of $r = 0.95$ ($CI_{95\%}$: 0.92 to 0.97) with our computational model, showing that people can continue to make quantitative knowledge inferences in contexts with real-world objects and variable adjective types. Critically, the correlation strength was comparable across all adjective types, with $r = 0.97$ ($CI_{95\%}$: 0.92 to 0.99) for color modification, $r = 0.95$ ($CI_{95\%}$: 0.85 to 0.98) for size modification, and $r = 0.96$ ($CI_{95\%}$: 0.87 to 0.98) for category modification. Moreover, participants' preferred identified referent matched our model's in 83.33% of trials ($n = 40$ of 48 trials; $P < 0.0001$ by binomial test with chance set to 0.25).

DISCUSSION

Our results show that people can make nuanced mental state attributions from the simplest linguistic events, such as the inclusion or omission of a single adjective in referential communication. In experiment 1, people were able to jointly infer speakers' intended referents and beliefs, and the confidence they reported in these inferences matched the quantitative structure of our theory of mind model. These results mirror the fine-grained inferences characteristic of "core" theory of mind tasks (*1*, *4*, *6*). Furthermore, in experiment 2, people showed how they adjust their mental state inferences

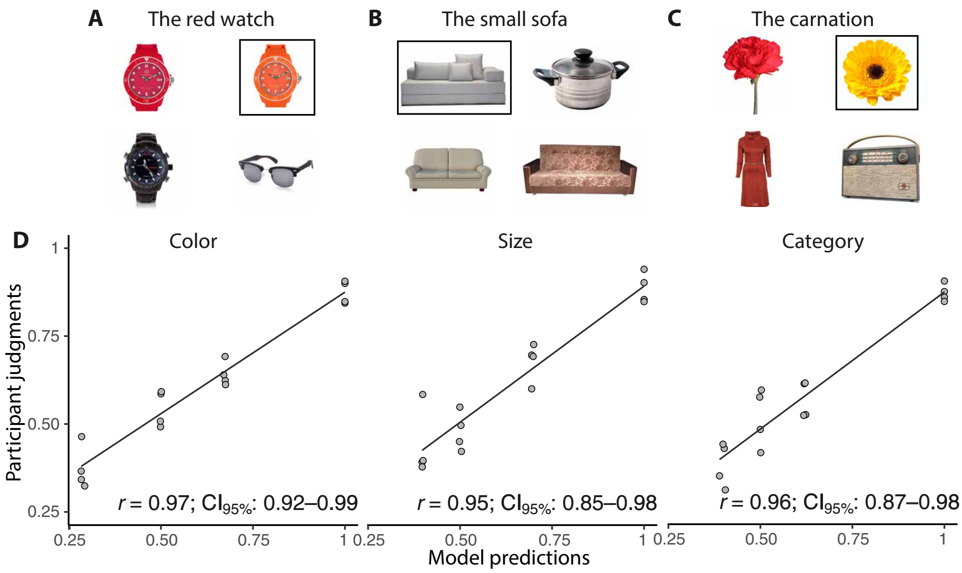

Fig. 5. Experiment 3 examples and overall results. (**A** to **C**) Three examples from the experiment using (A) color modification, (B) size modification, and (C) noun choice. In each trial, participants had to identify the referent and report their belief that the speaker could see the object surrounded by the black rectangle. (**D**) Experiment results. *x* axis shows model predictions, and *y* axis shows average participant judgments.

based on a brief exposure to speakers' propensity to use adjectives redundantly. Last, in experiment 3, people were able to jointly infer speakers' intended referents and beliefs based on how speakers referred to real-world objects using different types of adjectives and nouns with different degrees of specificity. Together, these experiments show that people can perform joint inferences about speakers' communicative intent and mental states from people's utterances at high levels of precision and accuracy.

While our goal was to characterize how people extract mental state information from language, experiment 1 also revealed an interesting finding: Reference can often be resolved accurately without attending to speakers' knowledge, as our deductive model also captured participants' pattern of reference resolution. This was only possible, however, because most of the referential expressions we used uniquely identified the intended target. Thus, the extent to which listeners can resolve reference without considering speakers' mental states may depend on speakers' willingness to ensure that they provide sufficient information for listeners to begin with. These findings suggest a trade-off between language production and language comprehension, where one interlocutor's usage of theory of mind may allow their communicative partner to rely less on theory of mind than would be otherwise necessary. These findings therefore inform a broader debate on how cognitive effort is divided between speakers and listeners to achieve successful communication (*45*, *48*, *49*).

The skilled use of theory of mind in communication that we identified here is all the more remarkable when we consider that speakers often use adjectives redundantly, especially color (*39*, *44*), and listeners must adjust their inferences accordingly. Given the general tendency to produce overinformative descriptions, one might have thought that listeners would not read too much into a speaker's choice of referential expression. However, our study shows that people can rely on subtle linguistic choices to derive quantitative theory of mind inferences that are sensitive to both speakers' choice of words and their propensity to use them redundantly.

At the same time, our work only established this capacity in contexts where interlocutors are explicitly asked to infer mental states with no time pressure. Can people also derive these inferences in real time? And, if so, do they do so in everyday conversation? Although these questions remain open, related research has found that listeners can make real-time inferences to anticipate what a speaker is talking about based on subtle speaker word choices like the ones that we manipulated here (*39*, *46*, *50*, *51*). These results show that listeners can derive online inferences about referential intent, but they leave open the question of whether this capacity extends to inferences about speaker knowledge. Recent research has also found that people spontaneously infer and track interlocutor knowledge in communicative interactions (*43*, *52*, *53*), confirming that theory of mind is also deployed in real-time communication. These findings suggest that listeners actively and spontaneously infer speakers' knowledge during communication. However, it is possible that these real-time inferences are coarse and qualitative (perhaps better approximated by the deductive model for computational convenience) and that nuanced mental state inferences require additional time and volition. Even if this is the case, however, these slower inferences could still be crucial in conversation and social interactions, enabling listeners to build high-resolution models of speakers' minds throughout an extended conversation. Therefore, having identified a highly skilled use of theory of mind in offline communication, future studies should investigate the degree of precision of mental state inferences in real-time communication, and whether different factors (such as speaker preferences or contextual relevance) may determine whether theory of mind inferences are coarse and qualitative or nuanced and quantitative.

Our work also opens a new question: What purpose do these nuanced mental state inferences serve? One possibility is that the precise and quantitative nature of mental state inference is a general signature of theory of mind. If so, then the pressure for quantitative inferences may have emerged from a pressure to understand physical action, but these inferences are also available in communication because language-based inferences rely on the same computations that underlie action-based inferences.

Such a view would suggest that mental state inferences operate over abstract representations that are modality independent. This idea is consistent with evidence that inferences from physical action are structured around an expectation that agents should expend as much effort as necessary to fulfill their physical goal, but no more (*5*, *54*), which parallels the expectation that agents should say as much as necessary to fulfill their communicative goal, but no more (*37*, *55*). Therefore, quantitative mental state inferences may depend on abstract representations of agents' effort relative to a baseline level of physical efficiency (*56*) and communicative efficiency (as shown in experiment 2).

Quantitative language-based inferences, however, may also be crucial for communicative success and not a simple side effect of action-based inferences. In communicative interactions, how we choose to respond may depend not only on coarse guesses about other people's mental states but also on precise and accurate estimates of what they may or may not know (and may or may not want to know). Moreover, the high-level mental state attributions that we make in communicative interactions (e.g., inferring what someone may have implied by what they said or left unsaid) likely rely first and foremost on rapid analyses of people's choice of words. These communicative demands would put pressure on getting that first layer of mental state inferences right, preventing small errors to cascade onto larger errors when making broader judgments about other people's minds. Our work lays groundwork toward addressing these questions, enabling studies that can test potential cascading errors that could occur in extended conversations as a function of the resolution of the mental state inferences that interlocutors make.

To conclude, studies on how we infer other people's mental states have typically focused on observable action. Our work shows how people can also extract mental state information from speakers' choice of words—even those that do not directly encode mentalistic information—with high fidelity. These results most directly show how people come to build nuanced and accurate representations of each other's minds. At the same time, our results also speak to the interaction between language and theory of mind.

Theories about how mental state reasoning and language understanding interact have typically focused on extreme cases, either advancing accounts that attempt to explain language understanding in terms of nonmentalistic processes (*57*, *58*) or using intrinsically mentalistic constructs as theoretical building blocks (*37*, *48*). Our work provides a different approach toward advancing this debate. By developing computational models of mental state inferences in language, we can shed light on how language relies on theory of mind, how often and how fast this interaction might take place, and how it varies across communicative situations and speakers. In turn, we may come to better understand how the neural circuitry behind these computations operates (*59*) and what happens when the interaction between language and theory of mind goes awry. Above and beyond all these questions, by charting the connections between how we infer the thoughts in other minds and how we share thoughts across minds, we will better understand what makes us exceptional social creatures, from passing interactions with strangers to extended conversations across our lifetimes.

MATERIALS AND METHODS

All research was approved by Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (MIT COUHES; "Development of Visual Perception," no. 0403000050R016) and Yale Institutional Review Board ("Online reasoning," no. 2000020357).

Experiment 1

Sixty participants (mean age, 35.22; range, 18 to 73) were recruited from Amazon's Mechanical Turk platform. Stimuli consisted of 28 trials. Each trial in turn consisted of two 2 × 2 grids with four colored geometrical shapes and a definite description of one of them. Stimuli were randomly split into two subsets of 14 trials, and participants were presented with only one of the two lists (n = 30 per condition; see the Supplementary Materials for details). In each trial, participants were told that the speaker intended to refer to only one of the corners and that the speaker could not see one of the four corners. Participants were also told that the speaker could not see the same corner in each display. The speaker's choice of referential expression to single out the target was written above each display. Trial order was randomized across participants. The pairs of displays were randomly ordered and rotated in each trial.

In each trial, participants were presented with three "trackpads," each with a circle that the participants could position anywhere they wished. Participants had to input their belief about the referent in each display in the first two trackpads and their beliefs about the blind spot in the third trackpad. Participants were given two examples of how to use the 2D trackpads and two examples of complete trials to show them how to reason about the blind spot by considering both displays (see the Supplementary Materials).

Model predictions

Our model outputs a posterior distribution over each of the four corners (i.e., four probabilities, one for each corner) for both each referent inference and for the belief inference (totaling 12 predictions per trial). To compare these predictions to participant judgments (three trackpad positions per trial), we transformed model predictions into trackpad positions by setting the x position to the sum of the two probabilities of the right corners and the y position to the sum of the two probabilities of the top two corners.

Experiment 2

One hundred forty-five participants (mean age, 36.08; range, 21 to 67) were recruited from Amazon's Mechanical Turk platform and randomly assigned to one of the four speaker conditions (see Results). Twenty-three of these participants were excluded from analysis for failing to correctly identify the referent two or more times during the first four unambiguous trials (including these participants in our analyses does not affect our conclusions; see the Supplementary Materials for details). Stimuli consisted of five displays of four geometrical shapes in each of the four conditions (see Fig. 4). The description of the target was written above each display. Displays were randomly rotated for each participant (thus counterbalancing the location of the referents and blind spots). The displays were presented one a time and remained on the screen in subsequent trials to avoid memory load.

Participants were assigned to one of four conditions that varied the speaker's propensity to use redundant size adjectives in trials 1 to 3. In the first condition, the speaker used no redundant size adjectives. In the second condition, the speaker used a redundant size adjective once (on trial 2). In the third condition, the speaker used a redundant size adjective twice (on trials 1 and 3). Last, in the fourth condition, the speaker used redundant adjectives on all three trials.

Participants were told that their partners in the game could only see three of the four shapes in each display, although they were unaware of this. In the first four trials, participants were asked to identify the target the speaker was referring to (using a 2D trackpad) and to report the speaker's overall propensity to speak redundantly (using a continuous slide bar with labels "never" and "always" on the two extremes, and "sometimes" in the middle). On the fifth trial, participants were asked the same two questions as before, and they were presented with an additional trackpad that asked them to identify the speaker's blind spot. Raw data visualized in Fig. 4 were obtained through kernel density estimation with bandwidth 0.75.

Model predictions

Model predictions were generated in the same way as in experiment 1. In addition, we also included estimates of the expected value of the speaker's degree of redundancy by computing the full joint distribution $p(t, b, r \mid u)$ rather than marginalizing the posterior over r, as shown in Eq. 1.

Experiment 3

Two hundred participants (mean age, 37.63; range, 21 to 74 were recruited from Amazon's Mechanical Turk platform ($N = 50$ per condition). Stimuli consisted of 48 displays, each consisting of a set of four pictures, a target, and a referent (see Fig. 5A). The description of the target was written above each display. Displays were randomly rotated for each participant (thus counterbalancing the location of the referents and blind spots).

Displays were evenly distributed among three conditions ($n = 16$ by condition): color condition (e.g., "Select the orange butterfly"), size condition (e.g., "Select the small sofa"), and category condition (understood as the specificity of the noun; e.g., "Select the flower" versus "Select the carnation"). For each condition, we considered four trial types (in decreasing order of certainty): (i) direct reference: the virtual partner refers directly to the target, which suggests that they can see it (e.g., "Select the dog," when the picture of the dog appears inside the frame); (ii) indirect reference: the virtual partner refers to the target indirectly by using an adjective contrastively, which suggests that they are likely to see the target (e.g., "the small sofa," when the big sofa appears inside the frame); (iii) contrastive reference: the virtual partner uses an adjective contrastively but the target is one of two contrast objects, which suggests that they may or may not see the target (e.g., "the orange butterfly," when the target is a blue butterfly but there is also a red butterfly in the display); and (iv) ambiguous reference: the virtual partner produces an ambiguous instruction, which suggests they cannot see the target (e.g., "the flower," when there is a carnation in the display and a buttercup appears inside the frame). Four lists of 12 trials were built by crossing the three properties of the referent that were manipulated (i.e., color, size, and specificity) with the four types of instructions (i.e., direct, indirect, contrastive, and ambiguous reference).

Participants were randomly allocated to one of the four lists of materials. The instructions explained that the partner who could see three of the objects in each display and may or may not see the fourth object inside a frame. Participants' task was twofold: They had to indicate which of the four objects the virtual partner was asking them to select (by clicking one of four radio buttons, each corresponding with a quadrant in the display), and they had to indicate how likely the virtual partner was to see the object inside a frame (by using a 0-to-10 scale, ranging from "definitely not" to "definitely yes," with the middle point indicating "maybe"). As part of the instructions, participants were shown a sample display that did not include an adjective in the instructions. Before they could start the task, they had to respond correctly to three questions to ensure that they had understood the instructions.

Model predictions

Model predictions were generated in the same way as in experiment 1, with the difference that the hypothesis space consisted of whether the speaker had full knowledge of the display or whether they were unaware of the target.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abj0970

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. A. Jern, C. G. Lucas, C. Kemp, People learn other people's preferences through inverse decision-making. *Cognition* **168**, 46–64 (2017).
2. A. Jern, C. G. Lucas, C. Kemp, Evaluating the inverse decision-making approach to preference learning, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2011), pp. 2276–2284.
3. C. L. Baker, R. Saxe, J. B. Tenenbaum, Action understanding as inverse planning. *Cognition* **113**, 329–349 (2009).
4. C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum, Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).
5. J. Jara-Ettinger, H. Gweon, L. E. Schulz, J. B. Tenenbaum, The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn. Sci.* **20**, 589–604 (2016).
6. J. Jara-Ettinger, L. E. Schulz, J. B. Tenenbaum, The naïve utility calculus as a unified, quantitative framework for action understanding. *Cogn. Psychol.* **123**, 101334 (2020).
7. J. Jara-Ettinger, F. Sun, L. Schulz, J. B. Tenenbaum, Sensitivity to the sampling process emerges from the principle of efficiency. *Cognit. Sci.* **42**, 270–286 (2018).
8. J. Jara-Ettinger, Theory of mind as inverse reinforcement learning. *Curr. Opin. Behav. Sci.* **29**, 105–110 (2019).
9. T. Ullman, C. Baker, O. Macindoe, O. Evans, N. Goodman, J. Tenenbaum, Help or hinder: Bayesian models of social goal inference, in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta, Eds. (Curran Associates, Inc., 2009), vol. 22, pp. 1874–1882.
10. N. Budwig, A developmental-functionalist approach to mental state talk, in *Language, Literacy, and Cognitive Development*, (Psychology Press, 2002), pp. 73–100.
11. B. Keysar, D. J. Barr, J. A. Balin, J. S. Brauner, Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychol. Sci.* **11**, 32–38 (2000).
12. S. Lin, B. Keysar, N. Epley, Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *J. Exp. Soc. Psychol.* **46**, 551–556 (2010).
13. W. S. Horton, B. Keysar, When do speakers take into account common ground? *Cognition* **59**, 91–117 (1996).
14. L. W. Lane, M. Groisman, V. S. Ferreira, Don't talk about pink elephants! Speakers' control over leaking private information during language production. *Psychol. Sci.* **17**, 273–277 (2006).
15. E. Kronmüller, T. Morisseau, I. Noveck, Show me the pragmatic contribution: A developmental investigation of contrastive inference. *J. Child Lang.* **41**, 985–1014 (2014).
16. J. E. Hanna, M. K. Tanenhaus, Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognit. Sci.* **28**, 105–115 (2004).
17. J. E. Hanna, M. K. Tanenhaus, J. C. Trueswell, The effects of common ground and perspective on domains of referential interpretation. *J. Memory Lang.* **49**, 43–61 (2003).
18. D. Heller, D. Grodner, M. K. Tanenhaus, The role of perspective in identifying domains of reference. *Cognition* **108**, 831–836 (2008).
19. K. P. Kording, D. M. Wolpert, Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
20. M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
21. D. Kersten, P. Mamassian, A. Yuille, Object perception as bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).

22. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).
23. G. Gigerenzer, P. M. Todd, *Simple Heuristics that Make Us Smart* (Oxford Univ. Press, 1999).
24. D. Grodner, J. Sedivy, The effect of speakerspecific information on pragmatic inferences, in *The Processing and Acquisition of Reference* (MIT Press, 2005).
25. A. Pogue, C. Kurumada, M. K. Tanenhaus, Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Front. Psychol.* **6**, 2035 (2016).
26. R. Ryskin, C. Kurumada, S. Brown-Schmidt, Information integration in modulation of pragmatic inferences during online language comprehension. *Cognit. Sci.* **43**, e12769 (2019).
27. M. C. Frank, N. D. Goodman, Predicting pragmatic reasoning in language games. *Science* **336**, 998–998 (2012).
28. N. D. Goodman, M. C. Frank, Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* **20**, 818–829 (2016).
29. C. Qing, M. Franke, Variations on a bayesian theme: Comparing bayesian models of referential reasoning, in *Bayesian Natural Language Semantics and Pragmatics* (Springer, 2015), pp. 201–220.
30. M. Franke, J. Degen, Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PLOS ONE* **11**, e0154854 (2016).
31. G. Scontras, N. D. Goodman, Resolving uncertainty in plural predication. *Cognition* **168**, 294–311 (2017).
32. N. D. Goodman, A. Stuhlmüller, Knowledge and implicature: Modeling language understanding as social cognition. *Topics Cogn. Sci.* **5**, 173–184 (2013).
33. P. Shafto, N. D. Goodman, T. L. Griffiths, A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cogn. Psychol.* **71**, 55–89 (2014).
34. J. Wang, P. Wang, P. Shafto, Sequential cooperative bayesian inference, in *Proceedings of the 37th International Conference on Machine Learning* (PMLR, 2020), pp. 10039–10049.
35. S. R. Searcy, P. Shafto, Cooperative inference: Features, objects, and collections. *Psychol. Rev.* **123**, 510–533 (2016).
36. N. D. Goodman, J. B. Tenenbaum, J. Feldman, T. L. Griffiths, A rational analysis of rule-based concept learning. *Cognit. Sci.* **32**, 108–154 (2008).
37. H. P. Grice, Logic and conversation, in *Speech Acts* (Brill, 1975), pp. 41–58.
38. P. Rubio-Fernandez, F. Mollica, J. Jara-Ettinger, Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *J. Exp. Psychol. Gen.* **150**, 583–594 (2021).
39. J. C. Sedivy, Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *J. Psycholinguist. Res.* **32**, 3–23 (2003).
40. P. E. Engelhardt, K. G. Bailey, F. Ferreira, Do speakers and listeners observe the Gricean maxim of quantity? *J. Mem. Lang.* **54**, 554–573 (2006).
41. P. E. Engelhardt, F. Ferreira, Do speakers articulate over-described modifiers differently from modifiers that are required by context? implications for models of reference production, Language. *Cogn. Neurosci.* **29**, 975–985 (2014).
42. A. S. Nadig, J. C. Sedivy, Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* **13**, 329–336 (2002).
43. P. Rubio-Fernández, The director task: A test of theory-of-mind use or selective attention? *Psychon. Bull. Rev.* **24**, 1121–1128 (2017).
44. P. Rubio-Fernandez, Overinformative speakers are cooperative: Revisiting the Gricean maxim of quantity. *Cognit. Sci.* **43**, e12797 (2019).
45. R. D. Hawkins, H. Gweon, N. D. Goodman, The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognit. Sci.* **45**, e12926 (2021).
46. J. C. Sedivy, 17 evaluating explanations for referential context effects: Evidence for cricean mechanisms in online language interpretation, in *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions* (MIT press, 2005), pp.345.
47. C. Kennedy, L. McNally, Scale structure, degree modification, and the semantics of gradable predicates. *Language* **81**, 345–381 (2005).
48. D. Sperber, D. Wilson, *Relevance: Communication and Cognition* (Harvard Univ. Press, 1986), vol. 142.
49. T. F. Jaeger, R. P. Levy, Speakers optimize information density through syntactic reduction, in *Advances in Neural Information Processing Systems* (MIT Press, 2006), pp. 849–856.
50. P. Rubio-Fernandez, J. Jara-Ettinger, Incrementality and efficiency shape pragmatics across languages. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 13399–13404 (2020).
51. P. Rubio-Fernandez, H. A. Terrasa, V. Shukla, J. Jara-Ettinger, Contrastive inferences are sensitive to informativity expectations, adjective semantics and visual salience. *PsyArXiv* 10.31234/osf.io/mr4ah (2019).
52. P. Rubio-Fernández, F. Mollica, M. O. Ali, E. Gibson, How do you know that? automatic belief inferences in passing conversation. *Cognition* **193**, 104011 (2019).
53. D. Heller, K. S. Gorman, M. K. Tanenhaus, To name or to describe: Shared knowledge affects referential form. *Topics Cognit. Sci.* **4**, 290–305 (2012).
54. G. Gergely, G. Csibra, Teleological reasoning in infancy: The naïve theory of rational action. *Trends Cogn. Sci.* **7**, 287–292 (2003).
55. D. Baldwin, J. Loucks, M. Sabbagh, *Pragmatics of Human Action* (Oxford Univ. Press, 2008).
56. J. D. Ongchoco, J. Jara-Ettinger, Beyond rationality: We infer other people's goals by learning agent-variable expectations of efficient action, paper presented at the Annual Meeting of the Cognitive Science Society, New Haven, CT, 2020.
57. C. Gauker, Zero tolerance for pragmatics. *Synthese* **165**, 359–371 (2008).
58. R. G. Millikan, *Language, Thought, and Other Biological Categories: New Foundations for Realism* (MIT Press, 1984).
59. A. M. Paunov, I. A. Blank, E. Fedorenko, Functionally distinct language and theory of mind networks are synchronized at rest and during language comprehension. *J. Neurophysiol.* **121**, 1244–1265 (2019).

Acknowledgments

Funding: This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by National Science Foundation (NSF) Science and Technology Center (STC) award CCF-1231216, and by the Research Council of Norway (Young Research Talent Grant awarded to P.R.-F.; award 230718). **Author contributions:** P.R.-F. and J.J.-E. conceptualized the study and experiments. P.R.-F. designed the stimuli for experiments 1 to 3, and J.J.-E. programmed the experiments. J.J.-E. implemented the computational model and analyzed the data in collaboration with P.R.-F. Both authors wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper, the Supplementary Materials, and/or at https://osf.io/h8qfy/.

Submitted 20 April 2021
Accepted 28 September 2021
Published 17 November 2021
10.1126/sciadv.abj0970

Quantitative mental state attributions in language understanding

Julian Jara-Ettinger and Paula Rubio-Fernandez

Sci. Adv. **7** (47), eabj0970. DOI: 10.1126/sciadv.abj0970

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.