Knowledge without belief

Bryan Gonzalez[1] Pauline Armary[2] James Dungan[3] Brent Strickland[2]
Joshua Knobe[4] Fiery Cushman[5] Jonathan Phillips[1,*]

7/2/2024

[1] Dartmouth College
[2] Intitut Jean Nicod
[3] Boston College
[4] Yale University
[5] Harvard University

[*] Correspondence: Jonathan Phillips <jonathan.s.phillips@dartmouth.edu>

Abstract: Five studies investigate the relationship between evaluations of knowledge and belief in theory of mind and demonstrate that evaluations of knowledge are made in the absence of evaluations of belief. Our studies find that (1) people can accurately evaluate others' knowledge before they evaluate their beliefs, (2) this pattern extends to participants with Autism Spectrum Disorder and cannot be explained by pragmatic differences in knowledge and belief ascriptions, (3) it also occurs cross-linguistically and is not accounted for by differences in word frequency, (4) it generalizes to the larger class of factive and non-factive attitudes (to which knowledge and belief respectively belong), and using fMRI data, (5) that the neural response that occurs when making evaluations of others' beliefs is absent when making similar evaluations of knowledge. Together, these studies demonstrate that human adults can attribute or deny knowledge states without first evaluating belief states. At a broad level, these findings suggest that knowledge representation is distinct from belief representation and offers a conceptually primitive way to represent others' minds.

Key words: knowledge; belief; theory of mind; autism; fMRI

Word count: 10094

Author note: The data, materials, and analysis code for all studies is publicly available at: https://doi.org/10.7910/DVN/IT0Z6S. The authors declare no conflicts of interest.

1 Knowledge without belief

This paper asks whether we can attribute knowledge without belief. Or, more precisely, it asks whether people can represent someone else as knowing something without representing them as believing that thing. There are theoretical and empirical reasons to favor both a positive and negative answer to this question. We make progress on this question across five studies, and they offer remarkably consistent answer using a combination of response time and fMRI data. Before turning to these results, it is worth explaining the motivation behind both the reasons for thinking there could not be knowledge without belief and the reasons for thinking there in fact could be.

1.1 Knowledge with belief

A standard way of thinking about the relationship between knowledge and belief across the cognitive sciences holds that the process of knowledge attribution is relativley complicated and may recruit or rely on more conceptually basic representations, like that of mere belief. It is not difficult to motivate such a perspective. For any case in which someone knows something, it seems intuitive that they must also believe it — after all, if you didn't even believe something in the first place, how could you possibly know it to be true? Perhaps then, attributing knowledge of something to others requires, at a minimum, representing them as believing that thing, and then additionally representing them as satisfying some additional criteria that are more specific to knowledge (e.g., that their belief is true, justified, formed through a reliable process, or what-have-you). On this hypothesis, when people are trying to determine whether someone knows something, they go through the usual process that would be associated with trying to determine whether someone believes something and also go through some additional processes to determine whether the person meets the various other conditions.

One sees aspects of this way understanding of the relationship between knowledge and belief both in standard philosophical analyses of knowledge (Ichikawa & Steup (2016)), and in recent philosophically informed empirical work on knowledge attribution (Buckwalter et al., 2015; Dudley et al., 2015; Dudley, 2018; Rose & Schaffer, 2013).

Additionally, this understanding of the relationship between knowledge and belief sits well with the work in developmental and comparative psychology that has argued for a 'core' or 'innate' capacity for theory of mind that essentially involves an ability for belief representation (Kovács et al., 2010; Leslie et al., 2004; Onishi & Baillargeon, 2005; Stich, 2013). Researchers aruging for such a capacity have sought to offer evidence for an ability for false belief representation in non-human primates (Hayashi et al., 2020; Krupenye et al., 2016) and in human infants as young as 7-months (Buttelmann et al., 2009; Kovács et al., 2010; Onishi & Baillargeon, 2005). If this picture of the core capacity for theory of mind is correct, and does essentially involves the ability to represent others' beliefs, then it would not be surprising if later-developing and more conceptually complicated mental state representations, e.g., 'imagining', 'being sure', or 'knowing' were built on top of conceptually primitive mental state representations, like mere belief, which are essential to the core capacity.

In short, this body of work collectively provides ample motivation for the view that representations of knowledge may depend on or involve prior representations of belief.

1.2 Knowledge without belief

Against this view, an alternative picture that has gained attention across the cognitive sciences is that the capacity to represent knowledge is more basic than the capacity to represent belief, and thus representations of knowledge do not depend on representations of belief (Phillips et al., 2020). Once again, it is not particularly difficult to see how such a view might have intuitive merit: many mundane cases of reasoning about someone's mind seem to concern others' knowledge without concerning their beliefs. For example, suppose you learn that the element with atomic number 3 is lithium, and you need to determine whether a particular chemistry professor knows that fact. One possible view would be that you can only figure this out by using the cognitive processes associated with belief attribution. (You have to figure out what the chemistry professor believes,

and also whether this belief satisfies some further conditions.) By contrast, another possibility would be that people can make use of distinct processes for knowledge attribution that are not the same as the processes for belief attribution. On this latter view, you might have an ability to very quickly determine that the chemistry professor has the relevant knowledge, and the process you use to make this determination might be completely separate from the one you would have used if you had instead been wondering about what the chemistry professor believes about the element with atomic number 3.

Some direct evidence for this second view comes from comparative research that has found that non-human primates often pass theory of mind tasks when they only involve representations of knowledge, but fail similar tests when they require belief representations (Horschler et al., 2019, 2020; Martin & Santos, 2014, 2016). Similarly, human infants are able to robustly pass theory of mind tasks that only require reasoning about others' knowledge long before they are able to pass tests that require belief representations (Luo & Johnson, 2009; Vouloumanos et al., 2014). Moreover, these findings concerning early-emerging knowledge representation have proven to be largely replicable, unlike the findings for early-emerging belief or implicit representations of belief (Holland & Phillips, 2020). Collectively, this research provides evidence for cases in which there is an ability to represent knowledge but no corresponding ability to represent belief. Obviously, in such cases, knowledge representation cannot depend on belief representation.

While these experimental tasks were designed to test for an underlying capacity for knowledge and belief representation, they do not explicitly employ the concepts of knowledge and belief. Importantly, however, other lines of research do explicitly employ these concepts, and find a remarkably similar pattern. For example, across languages, young children begin demonstrating a linguistic competence for communicating about knowledge (and related factive mental states, such as 'seeing') before they demonstrate a linguistic competence for communicating about belief (or related non-factive mental states, e.g., what others 'think') (Bartsch & Wellman, 1995; Harris et al., 2017; Wellman et al., 2001). Additionally, research in experimental philosophy on the ordinary concept of knowledge has provided clear cross-cultural evidence that there are cases in which people are willing to attribute knowledge but unwilling to attribute belief (Myers-Schulz & Schwitzgebel, 2013; Yuan & Kim, 2021). And finally, recent work using EEG has revealed that answering questions about what someone believes requires greater cognitive resources than answering matched questions about what someone knows, as measured in terms of P3b amplitude (Bricker, 2020). Clearly, these cases of explicit representations of knowledge are unlikely to depend on prior (or simultaneous) representations of belief.

1.3 Facitivity: A core difference between knowledge and belief

Perhaps the most fundamental way in which knowledge and belief differ is that knowledge is *factive* while belief is not. That is, knowledge is an attitude that one can only represent others as having toward truths (Lakoff, 1973; Williamson, 2002), while one can believe both things that are true and things that are not (Nagel, 2017; Phillips & Norby, 2019). This difference in knowledge and belief is not hard to see: If you don't think it's true that Al Gore invented the internet, then you can't represent someone else as knowing that he invented the internet; you can, of course, represent them as believing it. And you can also represent them as believing any other arbitrary proposition, e.g., that Reagan killed all the birds in 1986 and replaced them with spies who work for the bourgeoisie. Representations of belief are not constrained by the truth; representations of knowledge are. Thus, an important difference between factive and non-factive attitudes, in general, is that factive attitude representations may be simpler because the content represented is necessarily consistent with your own, and thus does not have to be represented separately from your own understanding of the content (Phillips & Norby, 2019). Given that this difference holds for factive and non-factive attitudes in general, it would obviously holds for knowledge and belief in particular. We take advantage of this broader difference in trying to understand the relationship between knowledge and belief.

1.4 Present studies

Five studies explored the possibility that people can attribute knowledge without attributing belief. The primary method we use is to measure response times for questions about whether a character in a scenario

does or does not know something vs. does or does not believe something. If participants are faster to correctly evaluate questions about knowledge than belief, we will have relatively clear evidence that people can attribute knowledge without first going through the process of attributing belief. Four studies collectively provide this evidence. In Experiment 1, we show that people can accurately evaluate others' knowledge before they evaluate their beliefs. In Experiment 2, we rule out the possibility that this response-time pattern is explained by pragmatic differences, and find that the patterns extends to participants with Autism. In Experiment 3, we demonstrate that the relative speed advantage of knowledge over belief ascriptions occurs cross-linguistically and is not accounted for by differences in word frequency. In Experiment 4, we find that this response-time difference generalizes to a larger class of factive and non-factive attitudes (to which knowledge and belief respectively belong). Finally, in Experiment 5, we move beyond response times and show that neural activity reveals a similar relationship between ascriptions of knowledge and belief. More specifically, we show that the neural response pattern that occurs when making evaluations of others' beliefs is absent when making similar evaluations of knowledge. Together, these studies demonstrate that human adults can attribute or deny knowledge states without prior or simultaneous evaluations of belief states. These findings collectively support the view that knowledge representations are a basic and distinct way in which we understand others' minds.

2 Experiment 1a-b

A simple way of determining whether people can make evaluations of knowledge in the absence of evaluations of belief is to investigate the speed with which these evaluations are made. If participants' evaluations of knowledge require or involve evaluations of belief, we would expect evaluations of knowledge ascriptions to be slower than evaluations of similar belief ascriptions. We began with a simple task that investigated the response times for evaluations of knowledge and belief ascriptions. Participants read about various people in a variety of different scenarios and then made truth value judgments of knowledge and belief ascriptions to those agents. In some cases, the knowledge and belief ascriptions were true, in others, they were false. We then compared the speed with which participants were able to correctly evaluate these mental state ascriptions.

2.1 Analysis approach

In all studies, response times for trials on which participants correctly assessed the truth of the knowledge and belief statements were analyzed with linear mixed effects models using the lme4 package in R (Bates et al., 2014), with both participants and scenario included as random factors. Participants were excluded from the analysis if they answered less than 67% of the questions correctly or if their mean response time was less than $1000ms$ or greater than $4000ms$. We additionally excluded all trials on which the response was given in less than $1000ms$ or longer than $4500ms$. We applied these same criteria to all of the response-time analyses in all of the experiments reported.

The data, materials, and analysis code for all studies is publicly available at: https://doi.org/10.7910/DVN/IT0Z6S. These studies and analyses were not preregistered. All studies were approved by either the Harvard University (IRB14-2016) or Dartmouth College (STUDY00032209; STUDY00032303) Institutional Review Boards. Informed consent was acquired from all participants.

2.2 Participants

Two hundred participants (M_{age}=32.76, SD_{age}=12.67; 108 females) were recruited through a psychology based website (http://www.moralsensetest.com). Experiment 1b was an exact replication with 501 participants recruited through Amazon Mechanical Turk.

2.3 Stimuli and procedure

Participants began by completing a demographic questionnaire, and two practice trials in which they were familiarized with the task they would be completing. Participants then completed twenty-four trials which involved reading a short vignette about an agent and then deciding whether a sentence about the story was true or false. Participants were instructed to indicate their responses as quickly as possible by pressing one of two keys on their keyboard. Twelve of these trials were 'distractor' trials in which participants were asked simple comprehension questions that did not mention the agent's mental states. These were included to prevent participants from anticipating the sentences they would be evaluating. In the remaining twelve trials, participants read a vignette that described the agent as either having true information about some proposition p (as in **True Information**), simply being ignorant of p (as in **No Information**), or believing some proposition q that was both false and inconsistent with p (as in **False Information**).

> **True Information**: Mira looks at the night sky with her telescope. She owns the most accurate books on the locations of the different planets throughout the year. Mira reads in her astronomy books that she can see Neptune through her telescope, and she waits until it's dark enough outside. She points her telescope towards the coordinates that her books specify for Neptune, and sees a bright dot in the middle of the sky. That bright dot is Neptune. She is excited that she found the planet she was looking for so easily.

> **No Information**: Mira likes looking at the night sky with her telescope. She owns the most accurate books on the locations of the different planets throughout the year. It is night and Mira decides not to read her astronomy books and instead just look through her telescope. Ignoring her book, she sets up her telescope and points it towards a group of dots that catch her attention. She looks into the telescope and she sees a bright dot in the middle of the sky. That bright dot is actually Neptune.

> **False Information**: Mira likes looking at the night sky with her telescope. She owns the most accurate books on the locations of the different planets throughout the year. It is night and Mira reads in her astronomy books that she can see Mercury through her telescope. Misreading her book, she sets up her telescope and points it towards the coordinates that her books specify for Neptune, which also happens to be in the sky. She looks into the telescope and she sees a bright dot in the middle of the sky. That bright dot is actually Neptune.

On each of the twelve test trials, participants were asked to evaluate the truth or falsity of a sentence about either the agent's knowledge, as in **Know**, or belief, as in **Think**. Critically, these sentences always concerned the proposition that was true:

> **Know**: Mira knows she is looking at Neptune.

> **Think**: Mira thinks she is looking at Neptune.[1]

Scenarios in all three conditions were constructed in such a way that the answer to both questions was true (i.e., True Information) or the answer to both questions was false (i.e., No Information and False Information), which allows us to directly compare response times across the two ascription types (Know vs. Think). We used Latin-square randomization to ensure that each participant saw all 12 distractor scenarios and all 12 test scenarios in random order, and judged two 'knows' ascriptions and two 'thinks' ascriptions for each of the three Information Conditions ('True Information' vs. 'No Information' vs. 'False Information').

[1] We used 'thinks' instead of 'believes' in these studies to better equate for word frequency and length.

2.4 Experiment 1a Results

The overall analysis of participants' response times revealed no main effect of Information Condition, $\chi^2(2) = 1.565$, $p = 0.457$, and no Information Condition × Ascription Type interaction, $\chi^2(2) = 1.602$, $p = 0.449$.[2] However, there was a significant main effect of Ascription Type, $\chi^2(1) = 20.057$, $p < 0.001$), such that participants were faster to correctly assess the truth of statements about whether the agent knows something ($M_{\text{rt}} = 2550.46$ms, $SD_{\text{rt}} = 739.06$) than statements about whether the agent thinks something ($M_{\text{rt}} = 2655.43$ms, $SD_{\text{rt}} = 751.27$).

2.5 Experiment 1b Results

Experiment 1b was a direct well-powered replication of Experiment 1a, with the only difference being that participant recruitment was through Amazon's Mechanical Turk. The analysis of participants' response times again revealed no main effect of Information Condition, $\chi^2(2) = 4.126$, $p = 0.127$, but did reveal a main effect of Ascription Type, $\chi^2(1) = 27.112$, $p < 0.001$), such that participants were faster to correctly assess the truth of statements about whether the agent knows something ($M_{\text{rt}} = 2459.37$ms, $SD_{\text{rt}} = 739.02$) than statements about whether the agent thinks something ($M_{\text{rt}} = 2517.93$ms, $SD_{\text{rt}} = 731.17$). Additionally, there was a significant Information Condition × Ascription Type interaction, $\chi^2(2) = 12.851$, $p < 0.002$. Given the identical methods of Study 1a and 1b, we next combined the data from both experiments to ask whether the interaction was significant across all of the data, and to provide the best estimate of the other effects.

2.6 Combined Analyses

The combined analysis continued to reveal no main effect of Information Condition, $\chi^2(2) = 2.68$, $p = 0.262$, but did reveal a main effect of Ascription Type, $\chi^2(1) = 47.013$, $p < 0.001$), such that participants were faster to correctly assess the truth of statements about whether the agent knows something ($M_{\text{rt}} = 2485.33$ms, $SD_{\text{rt}} = 740.05$) than statements about whether the agent thinks something ($M_{\text{rt}} = 2556.11$ms, $SD_{\text{rt}} = 739.23$). Moreover, we continued to observe a significant Information Condition × Ascription Type interaction, $\chi^2(2) = 12.385$, $p = 0.002$.

Planned pairwise comparisons were carried out using the Estimated Marginal Means package in R (Lenth, 2023). These analyses revealed that the largest difference in response times was found in the No Information conditions, $t(476.93) = -6.839$, $p < 0.0001$, followed by a somewhat smaller difference in the True Information conditions, $t(472.82) = -3.421$, $p = 0.009$. The smallest response time difference was found in the False Information conditions, where there was no significant difference between knowledge and belief ascriptions $t(475.64) = -2.287$, $p = 0.201$. Response time differences in each information condition showed the same pattern such that participants were faster to evaluate knowledge ascriptions than belief ascriptions (see Figure 1).

[2] The fixed and random effects structure for the full model was specified as: response.time ∼ belief.condition ∗ ascription.type + (1|scenario) + (belief.condition ∗ ascription.type|subj). We were not able to include random slopes for the scenario because the crossed nature of the random effects in our experiment prevented convergence. We employ a similar analysis approach throughout, except where explicitly noted.

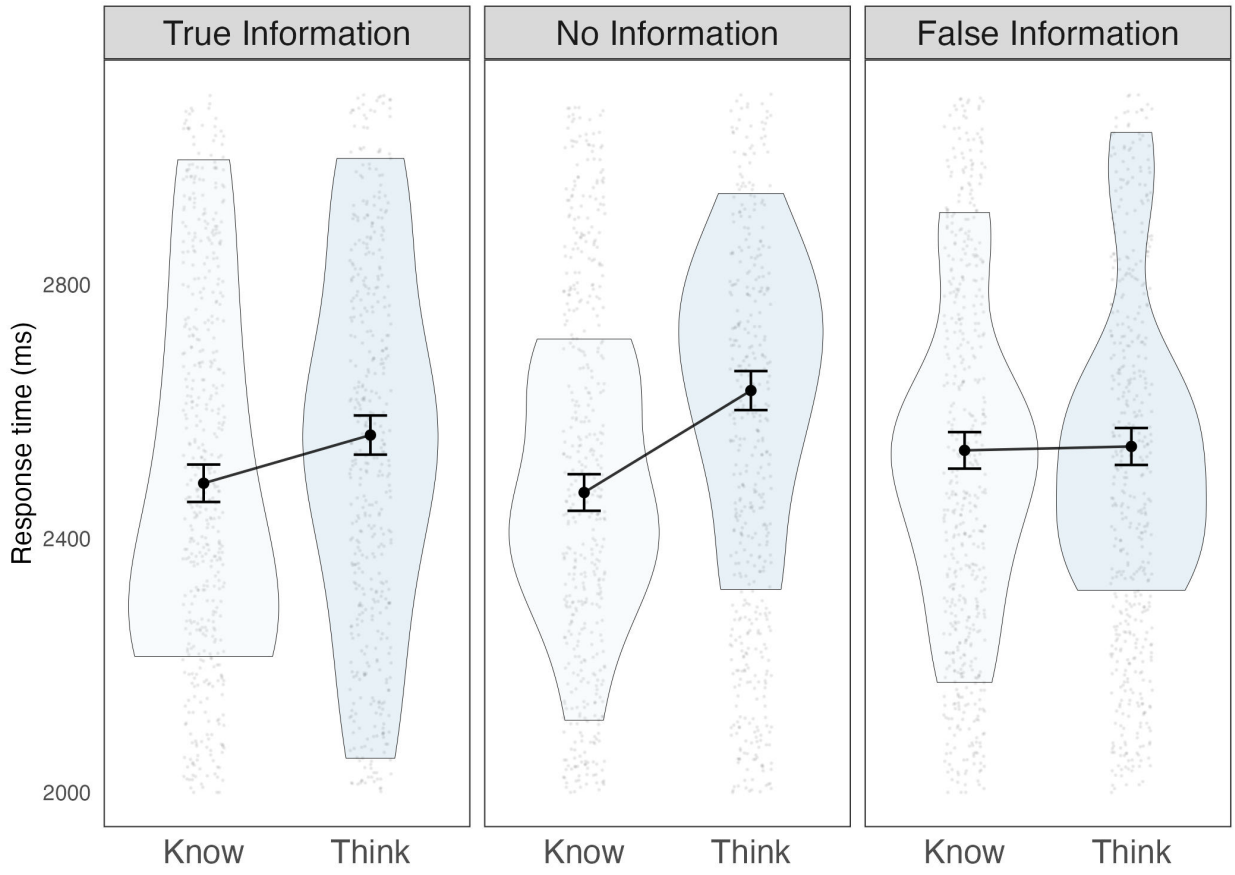

Figure 1. Response times for correct evaluations of knowledge ascriptions (left points and violins) and belief ascriptions (right points and violins) as a function of Information Condition (separate panels). Small points indicate trial-level responses, violins illustrate scenario- or item-level distributions, large dark points depict overall means, and error bars depict +/- 1 SEM.

2.7 Discussion

If evaluations of others' knowledge involve or require evaluations of belief, then we would expect that correctly determining whether someone knows something would take at least as long as determining whether someone believes that thing. We tested this prediction but found clear evidence that the opposite is true: correct evaluations of knowledge are made *before* correct evaluations of belief. In other words, the evidence suggests that participants are both attributing knowledge (in the True Information condition) and denying knowledge (in the No Information conditions) before they have been able to determine whether the agent believes the relevant claim.

A further question is why we observed an interaction effect, such that the difference in response times in assessing knowledge and belief ascriptions differed in the various information conditions. Before making much of this unpredicted interaction effect, we first want to ask whether it replicates, and ensure that the observed effects are not arising from some simpler, alternative explanation. One potential concern with this initial study is that it involves cases in which some of the belief ascriptions may be pragmatically odd Specifically, in cases where an agent meets all of the criteria for knowledge (as in the True Information example), it may sound unnatural to describe the agent as believing some proposition (as in **Think**) rather than describing the agent as knowing that proposition (as in **Know**). After all, the agent does not merely think the proposition in question, but also knows it. In such cases, while belief ascriptions will, of course, be strictly true, they may also be pragmatically odd or infelicitous. To illustrate with an example, imagine that there is an sandwich in the fridge and Sarah knows that fact. In such a case, it might seem odd to

use the sentence "Sarah thinks that there is an sandwich in the fridge." This sentence would be true, but all the same, people might think it would be more natural to use the sentence "Sarah knows that there is a sandwich in the fridge." In line with this intuition, work in linguistics has offered theories that may help explain the pragmatic infelicity of using the word "think" in cases like this one where "know" would also be true Heim (1991); Hirschberg (1985).

3 Experiment 2

In this Experiment, we aimed to test for the impact of pragmatic infelicity while also asking whether the observed effect generalizes to participants with Autism Spectrum Disorder (ASD). To ask whether differences in the pragmatic understanding of the sentences participants evaluated might explain the observed response time difference between knowledge and belief ascriptions, we measured the extent to which the sentences participants evaluated were perceived as pragmatically odd by participants. While pragmatic theories do not strictly predict a difference in the felicity of knowledge and belief ascriptions when the agent is ignorant or has a false belief (as both ascriptions in these cases are simply false), we decided to collect felicity judgments for all of the sentences used in the previous experiment.

Additionally, we also investigated whether the previously observed response time effect generalizes to a sample of participants with ASD. Prior research has demonstrated that individuals with autism present impairments in understanding pragmatic implicatures, especially when they involve reasoning about others' mental states (Kissine, 2012). Additionally, a separate well-established line of research on autism has demonstrated impaired mentalizing abilities as assessed through classic false-belief paradigms (Baron-Cohen et al., 1985). Accordingly, participants with ASD make an ideal population for exploring the previously observed response time difference: If pragmatic differences are responsible for the comparative delay observed in correctly evaluating belief ascriptions, we should expect differences in pragmatic understanding to be reflected in the response time patterns between the ASD group and the neurotypical control group. Alternatively, if the pattern of response times found in Experiment 1 instead resulted from knowledge assessments being made in the absence of calculations of others beliefs, we expect participants with autism to demonstrate the same, or more extreme differences in response times between knowledge and belief ascriptions, as they may specifically have difficulty with belief representation but not knowledge representations (Deschrijver & Palmer, 2020; Phillips et al., 2020).

3.1 Participants

Inclusion criteria required participants to be adults, fluent in English, and complete the experiment on a personal computer or laptop with a standard keyboard. The 611 participants included in subsequent analyses each had of a mean accuracy $\geq 67\%$ and a mean response time $> 1000ms$ and $< 4000ms$. 389 participants were recruited from Prolific (app.prolific.co) to the neurotypical group, ($M_{\text{age}} = 37.57$, SD_{age}=13.12, 50.9% female, 45% male, 3.34% other, 0.77% not disclosed).

Crowdsourcing marketplaces such as Amazon's Mechanical Turk or Prolific do not provide stratified participant pools based on an autism diagnosis. Therefore, recruitment of participants in this group came from study advertisements posted in various online forums and websites of organizations dedicated to the autism community during the period from August-October 2022. Given this open recruitment strategy, advanced security measures were taken to ensure that data collected from this group were provided by earnest participants[3] (Dennis et al., 2020). A short screener required participants to endorse eligibility criteria, including

[3]After screening, eligible participants received an automated email invitation to participate containing a unique study access link that could only be used once. The experiment was served to eligible participants from a custom-built encrypted server (https://www.linode.com/), secured with a firewall and other security precautions. Connections to the experiment server were tested to ensure that the experiment was not accessed through a virtual privacy network, proxy, relay network, or tor node. Experiment access was also restricted to connections from English-speaking countries according to ISO-639 standards. Finally, two unrelated open-ended free text response questions were required items in the experiment and manually screened for nonsensical or suspicious responses.

fluency in English, being at least 18 years of age, a computer or laptop to complete the experiment (no phones/tablets, etc.), and self report of a clinical autism diagnosis. Overall, 222 participants recruited to the autism group met the same accuracy and mean response time criteria described above for neurotypicals, (M_{age} = 28.12, SD_{age}=5.15, 43.24% female, 51% male, 4.95% other, 0.45% not disclosed). Importantly, participants in the autism group scored significantly higher on the 10-item Autism Quotient scale ($M_{\text{AQ-10}}$ = 5.74, $SD_{\text{AQ-10}}$ = 2.7) than participants in the neurotypical group ($M_{\text{AQ-10}}$ = 3.72, $SD_{\text{AQ-10}}$= 2.26), $t(397.7)$ = -9.43, $p < 0.0001$.

3.2 Stimuli and procedures

The experiment was conducted in two blocks. The first consisted of the same stimuli and procedures as Experiment 1a-b, in which the 3 (Information Condition: No Information, True Information, False Information) × 2 (Ascription Type: Know, Think) design conditions were randomized and counterbalanced across 12 unique vignette contexts. 12 distractor trials, in which participants were asked to evaluate the veracity of simple facts about a vignette, were also presented, but excluded from all analyses. Response times were recorded while participants evaluated mental state ascriptions as true or false. The second block consisted of the same conditions, such that the 12 vignettes and mental state ascriptions in the first block were presented a second time in randomized order. Each participant completed a training session in which the felicity judgment task was thoroughly explained. In addition, they completed two practice trials using statements that were clearly felicitous or clearly infelicitous, and were given feedback on their responses. On the test trials, participants were asked to judge how much the target statement ascribing a mental state to the agent in the vignette seemed infelicitous, and responded on a Likert scale from 1 ("sounds very normal") to 7 ("sounds very weird"). Following completion of the first two blocks, participants completed a 10-item Autism Quotient scale, provided demographic information, and were debriefed.

3.3 Results

3.3.1 Response-time analysis

We first asked whether we observed similar response-time patterns observed in Study 1 across both neurotypical and autistic participants. To do so, we began by analyzing the response time data to investigate whether there was a three-way interaction between participant group (neurotypical vs. autistic), Information Condition and Ascription Type, and found that there was $\chi^2(2) = 6.525$, $p = 0.038$. To further investigate this interaction, we next separately analyzed the response-time data from the neurotypical and autistic participants.

3.3.1.1 Neuroptypical group Focusing first on the neurotypical participants, we found that we closely replicated the patterns observed in Study 1. Specifically, we found no main effect of Information Condition, $\chi^2(2) = 1.419$, $p = 0.492$, but again found a main effect of ascription type, $\chi^2(2) = 15.007$, $p < 0.001$, and also observed a significant interaction between Information Condition and Ascription Type, $\chi^2(2) = 28.624$, $p < 0.001$ (see Figure 2A). Planned pairwise comparisons using the Estimated Marginal Means (Lenth, 2023) revealed that the largest difference in response times was again found in the No Information conditions, *z.ratio* = -7.289, $p < 0.0001$, followed by a smaller difference in the True Information conditions, *z.ratio* = -2.199, $p = 0.028$, and in the False Information conditions, *z.ratio* = -2.503, $p = 0.012$.

3.3.1.2 Autistic group Focusing next on the autistic participants, we observed a variation on this broad pattern. We again found no main effect of Information Condition, $\chi^2(2) = 2.412$, $p = 0.299$, but more importantly did again observe a main effect of Ascription Type, $\chi^2(2) = 14.137$, $p < 0.001$. Additionally, however, we no longer observed a significant interaction between Information Condition and Ascription Type, $\chi^2(2) = 0.317$, $p = 0.854$. In contrast to the neurotypical group, we found that participants with autism

made judgments of knowledge faster than judgments of belief to a roughly similar extent across Information Conditions (see Figure 2B).

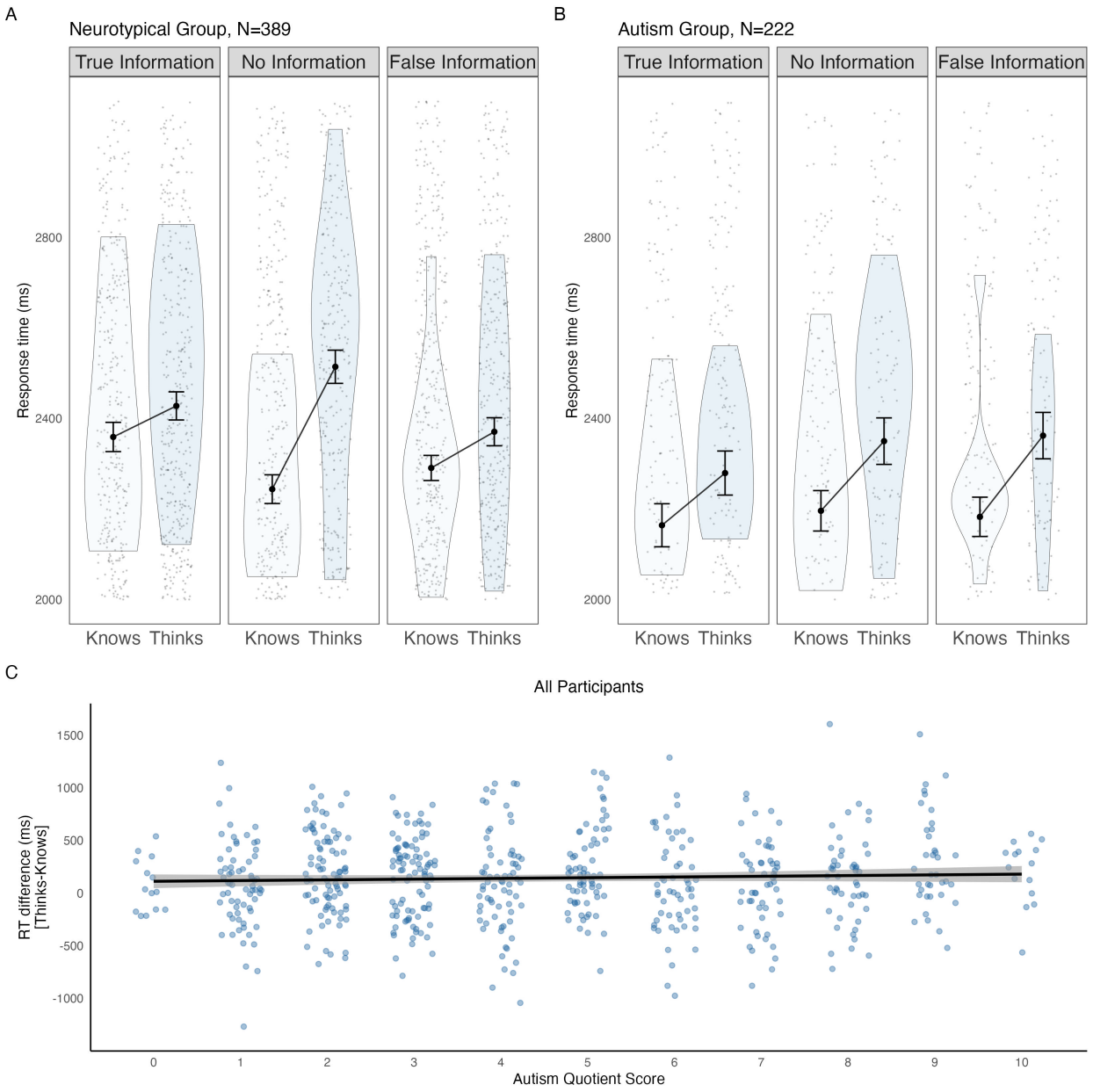

Figure 2. Response times of neruotypical (**A**) and autistic (**B**) groups for correct evaluations of knowledge ascriptions (left points and violins) and belief ascriptions (right points and violins) as a function of Information Condition (separate panels). Small points indicate trial-level responses, violins illustrate scenario- or item-level distributions, large dark points depict overall means, and error bars depict +/- 1 SEM. **C**: Difference in participant mean response time between belief and knowledge attribution trials as a function of participants' score on the Autism Quotient-10 scale across all participants.

3.3.1.3 Autism Quotient scale and response times In addition to looking for group-level differences between neurotypical and autistic participants, we also investigated the parametric relationship between autism scores (on the AQ-10) and response times in assessing mental states. To do so, we computed the difference between each subject's average response time for belief and knowledge ascriptions, such that more positive values indicate that knowledge ascriptions were faster than belief ascriptions while more negative

values indicate that belief ascriptions were faster. We then asked whether self-reported autism symptoms, as measured in all participants by the AQ-10, predicted the difference in speed with which knowledge and belief ascriptions are correctly evaluated. We did not find a significant relationship, and in fact, to the extent there is a systematic pattern, higher AQ-10 scores are associated with a *bigger* difference in response times, $F(1, 609) = 1.12$, $p = 0.29$ (Figure 2C).

3.3.2 Infelicity ratings

We next turned to analyzing participants' infelicity judgments to ask whether the perceived infelicity of the knowledge and belief ascriptions may be able explain the response-time patterns we observed. We first analyzed the overall patterns of infelicity judgments similar to the way in which we analyzed the response-time data. We found no significant three-way interaction between Participant Group, Information Condition, and Ascription Type, $\chi^2(2) = 0.298$, $p = 0.862$, and no Participant Group by Ascription Type interaction effect on infelicity judgments, $\chi^2(1) = 0.266$, $p = 0.606$.

However, we did observe a significant Participant Group by Information Condition interaction $\chi^2(2) = 175.55$, $p < 0.0001$. As can be seen by comparing Figure 3A and 3B, this interaction effect arose largely because participants in the autism group judged false mental state ascriptions of both knowledge and beliefs as relatively infelicitous, while participants in the neurotypical group did not. This pattern was significant for both the No Information condition ($t(620.56) = 14.33$, $p < 0.0001$), and the False Information condition ($t(620.06) = 16.4$, $p < 0.0001$). By contrast, there was no significant difference between these groups when the agent was described as having True Information ($t(623.21) = -1.23$, $p = 0.82$). In short, while we did find quite substantial differences between the infelicity judgments of neurotypical and autistic participants, these differences tracked the truth value of the mental state ascription but not the kind of mental state attributed (knowledge vs. belief).

Across both neurotypical and autistic groups, we also observed observed a significant Information Condition $\times$ Ascription Type interaction, $\chi^2(2) = 25.398$, $p < 0.0001$, which was roughly in line with the predictions of pragmatic theories (Heim, 1991; Hirschberg, 1985). As can be seen in Figure 3A-B, when agents have True Information, participants in both groups found it more natural to describe the agent as knowing that information ($M_{\text{infelicity}}$=1.58, $SD_{\text{infelicity}}$=1.36), than as thinking that information ($M_{\text{infelicity}}$=1.88, $SD_{\text{infelicity}}$=1.62), $t(68.29) = -4.28$, $p < 0.0001$. Critically however, across groups, this difference disappears when agents are described as having false information, $t(73.96) = 1.86$, $p = 0.434$, or no information, $t(84.44) = 1.31$, $p = 0.781$. The fact that the autism group also shows this particular pattern suggests that the relevant pragmatic implicature is likely highly conventionalized (Svindt & Surányi, 2021).

In sum, we observed two key patterns in participants' felicity judgments. The first is that we found group differences in felicity judgments whereby, whenever a knowledge or belief ascription was false, autistic participants found the ascriptions to sound worse than neurotypical participants. The second is that for the True Information condition, belief ascriptions sounded worse than knowledge ascriptions, while for the No and False Information conditions, this pattern disappeared. At a high level, these qualitative patterns make it straightforwardly unlikely that the previously observed response time differences (whereby knowledge is consistently faster than belief) could be explained by differences in the perceived infelicity of the ascriptions (compare the patterns in Figures 2A-B to those in Figures 3A-B).

Finally, to quantitatively verify these qualitative observations, we performed two additional analyses. First, we simply looked at the linear relationship between felicity judgments and response times for both the neurotypical group (Figure 3C), and the autistic group (Figure 3D). In the neurotypical group, we found a significant, albeit weak, relationship ($r = 0.06$, $p = 0.0004$). However, we found no significant relationship in the autistic group, $r = -0.04$, $p = 0.107$. Second, we asked whether, controlling for the felicity of each sentence as rated by each participant, we continued to observe the overall pattern whereby correct assessments of knowledge ascriptions were faster than correct assessments of belief ascriptions. We found that it was for both the neurotypical group, $\chi^2(2) = 14.099$, $p < 0.001$, and for the autism group, $\chi^2(2) = 14.099$, $p < 0.001$.

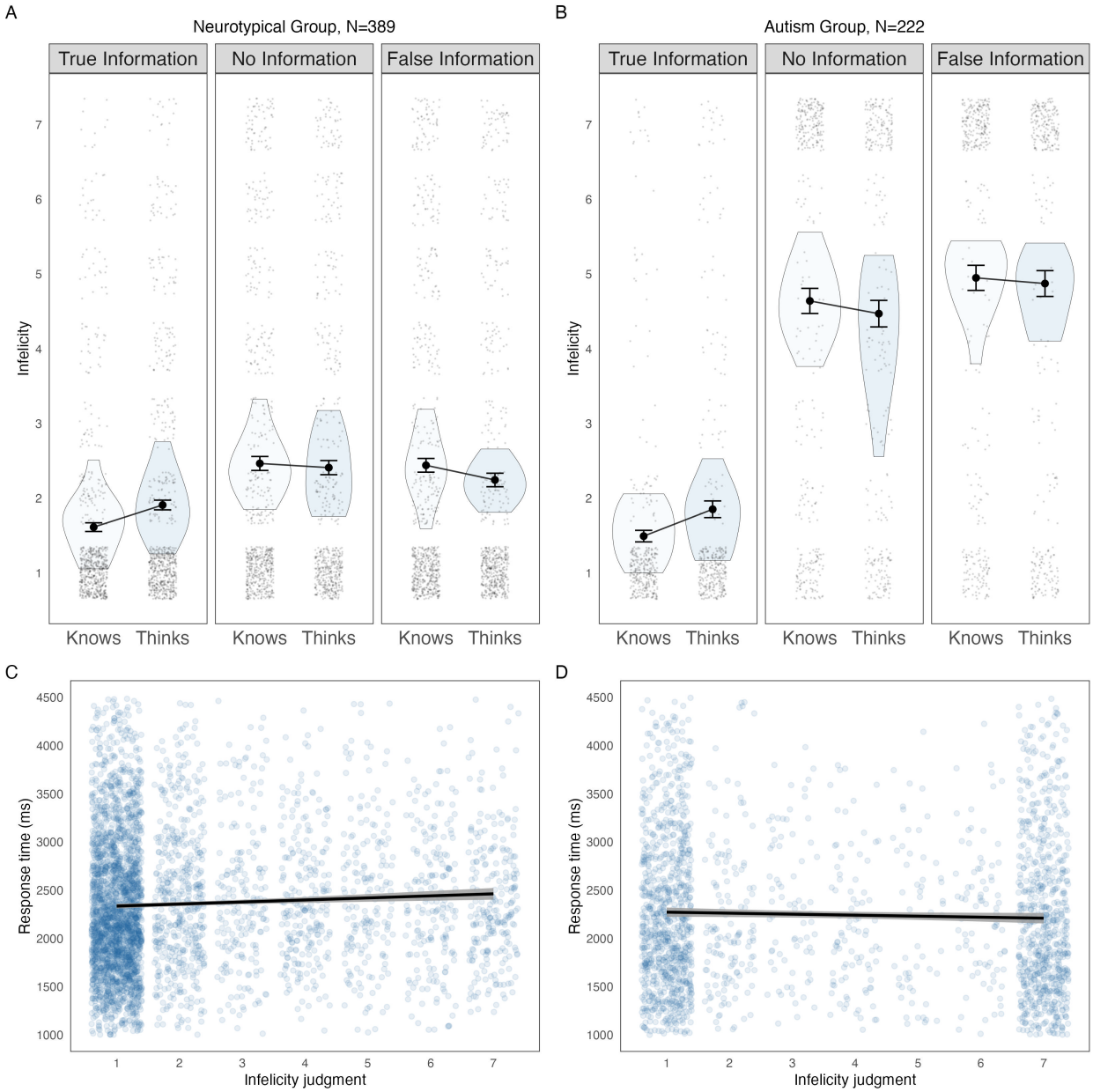

Figure 3. **A-B**: Mean infelicity ratings from neurotypical (**A**) and autsitic group (**B.**) for knowledge ascriptions (left points and violins) and belief ascriptions (right points and violins) as a function of Information Condition (separate panels). Small points indicate trial-level responses, violins illustrate scenario- or item-level distributions, large dark points depict overall means, and error bars depict +/- 1 SEM. **C-D**: Response times of neurotypical (**C**) and autism (**D**) groups for correct evaluations of mental state ascriptions (knowledge & belief) as a function of participants' infelicity judgments.

3.4 Discussion

In this study, we replicated Experiment 1 with a group of neurotypical participants and a group of autistic participants, and additionally asked all participants to evaluate the felicity of the mental state ascriptions. The analysis of the response-time data revealed that we not only replicated the finding that neurotypical participants were able to correctly evaluate knowledge ascriptions faster than they were able to correctly evaluate belief ascriptions, but that this pattern extended to participants with autism. We additionally asked

whether this response-time pattern could be explained by differences in the perceived felicity of knowledge and belief ascriptions, but found this explanation lacking on both a quantitative and qualitative level. Taken together, these results suggest that the response time patterns found in Experiments 1 and 2 do not arise from pragmatic differences in knowledge vs. belief ascriptions, and robustly generalize across samples from neurotypical populations and populations with autism.

4 Experiment 3

We next asked whether the observed response time difference between 'know' and 'think' generalized to languages other than English. To do this, a highly similar experiment was conducted in French using the mental state verbs 'savoir' and 'penser' instead of 'know' and 'think' (respectively). In addition to providing a general test of whether the observed effect could have arisen from idiosyncratic features of the English used in the prior experiment, French provides a particularly strong test because, unlike English, the French term used for belief ascriptions 'penser que' is ≈ 1.49 times more frequent than French term used for knowledge ascriptions 'savoir que.'[4] Thus, if the difference in response times merely reflected lexical frequency, we should expect evaluations of beliefs to be faster than evaluations of knowledge in French, which is the opposite of the prediction of our more general theory of mind based account.

4.1 Participants

150 participants (M_{age}=37.7, SD_{age}=12.16; 83 females) were recruited and paid €0.75 through Foule Factory (https://www.foulefactory.com/).

4.2 Stimuli and procedure

The methods and procedures in this experiment were similar to that of Experiment 1a-b, except that the study was translated into French, and the English names were replaced with more standard French names. Thus, for example, rather than the **Know** and **Think** example sentences in Experiment 1, participants evaluated **Savoir** and **Penser** as below.

Savoir: Nora sait qu'elle regarde Neptune.

Penser: Nora pense qu'elle regarde Neptune.

4.3 Results

As in Experiment 1 and 2, data were excluded at the participant- and trial-level, and then analyzed using linear mixed-effects models. We found no main effect of Information Condition, $\chi^2(2) = 4.022$, $p = 0.134$. However, there was again a highly significant main effect of Ascription Type, $\chi^2(1) = 22.246$, $p < 0.001$, such that participants were faster to correctly assess the truth of statements about what the agent knows ($M =$ 2546.8ms, $SD = 719.03$) than statements about what the agent thinks ($M = 2701.13$ms, $SD = 716.9$). Once again, there was also a significant Information Condition × Ascription Type interaction, $\chi^2(2) = 7.434$, $p =$ 0.024 (see Figure 5).

[4]Lexical frequency was computed using Google NGram using data from 2007-2008, which was the most recent year available at the time of calculation. For both French and English, we summed the frequency of the following forms of 'Savior' and 'Penser' or 'Know' and 'Think': infinitival, first person singular, second person singular, third person masculine, third person feminine, first person plural, second-person plural (French only), third-person plural (English), and masuline and feminine third person plural (French). We then divided the summed frequency of 'know' or 'savoir' by the summed frequency of 'think' or 'penser' respectively.

Planned pairwise comparisons revealed the Information Condition x Ascription Type interaction was driven by a significant difference in participants' response times evaluating knowledge and belief in the No Information condition, $t(93.1845205) = -5.006$, $p = 0$. Smaller, non-significant, differences were found in the False Information, $t(93.5183022) = -1.81$, $p = 0.464$, and True Information $t(94.2469764) = -1.597$, $p = 0.603$, conditions.

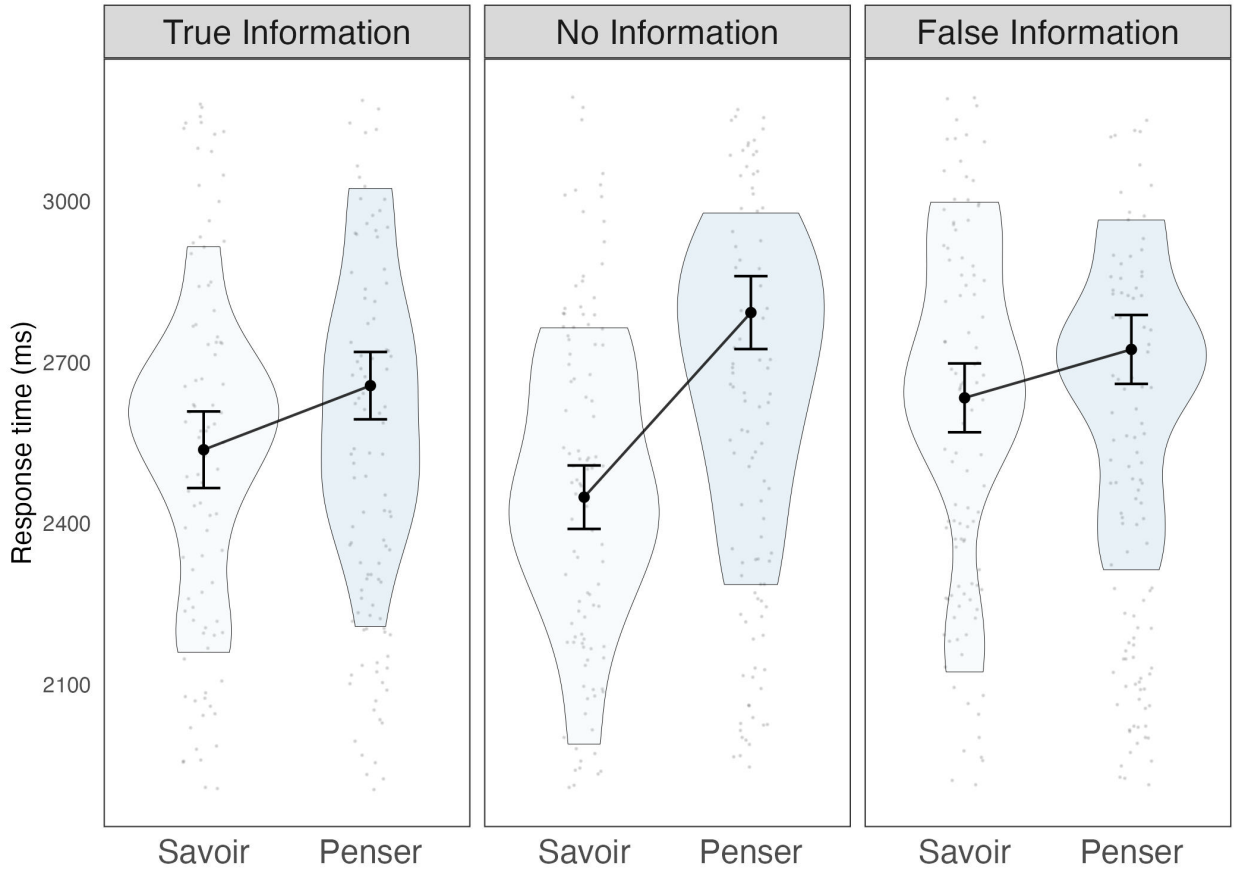

Figure 5. Response times for correct evaluations of knowledge ('Savoir') ascriptions (left points and violins) and belief ('Penser') ascriptions (right points and violins) as a function of Information Condition (separate panels). Error bars depict Small points indicate trial-level responses, violins illustrate scenario- or item-level distributions, large dark points depict overall means, and error bars depict +/- 1 SEM.

4.4 Discussion

This third experiment provides cross-linguistic evidence that evaluations of knowledge ascriptions occur prior to equivalent belief ascriptions, and thus that evaluations of others' knowledge are made in the absence of similar evaluations of their beliefs. In addition, it provides clear evidence against an explanation of this pattern in terms of lexical frequency.

An important remaining question is whether the observed difference is specific to knowledge and belief or whether it reflects a more general difference between different categories of mental state verbs of which knowledge and belief are a part. We pursue this possibility in Experiment 4.

5 Experiment 4

Knowledge and belief are just two representative examples of the broader classes of factive and non-factive mental state attitudes. Similar to knowledge, you cannot represent someone as seeing or hearing something that is false, because these attitudes are factive. And similar to belief, you *can* represent someone as assuming or guessing things you know are false, because these attitudes are non-factive. Thus, a sentence like "Claudia realized that John was in the building" implies that John actually was in the building, whereas a sentence like "Claudia assumed that John was in the building" does not imply that John actually was in the building. Accordingly, one possibility is that people will be generally faster to evaluate factive attitude ascriptions than non-factive attitude ascriptions. A plausible reason to predict this general difference is that factive mental state representations may be simpler because the content represented is necessarily consistent with one's own understanding, and thus does not have to be represented separately (Phillips & Norby, 2019). If this prediction proves true, it could point toward a more general explanation of why people are faster to evaluate knowledge ascriptions than belief ascriptions.

To clarify, our prediction is not that every factive attitude will be evaluated faster than every non-factive attitude; after all, there are many factors that will jointly determine evaluation time. Our prediction is instead that–other things being equal–it will take longer to evaluate the truth of a given non-factive attitude than an otherwise similar factive attitude. We next investigate this possibility by considering the speed with which participants evaluate a larger range of factive and non-factive mental state ascriptions.

5.1 Participants

250 participants (M_{age}=33.43, SD_{age}=9.32; 126 females) were recruited and paid through Amazon Mechanical Turk.

5.2 Stimuli and procedure

The methods and procedures in this experiment were similar to the preceding studies except that the term 'know' in the mental state ascription was replaced by one of a set of factive attitude verbs ('realize', 'is aware', 'understand', 'recognize'), and the term 'think' was replaced by a set of non-factive attitude verbs ('believe', 'guess', 'assume', 'imagine'). Thus, for example, instead of evaluating the truth or falsity of **knows** or **thinks** as Experiment 1, participants may have evaluated the truth or falsity of **factive** or **non-factive**, respectively.

> **Factive**: Mira recognizes that she is looking at Neptune.
>
> **Non-factive**: Mira believes that she is looking at Neptune.

Critically, these factive and non-factive terms were chosen such that the non-factive terms were both shorter in length and more frequent in use than the factive terms (Table 1), so that the predicted difference in response times (factive < non-factive) could not arise from differences in length or lexical frequency. Because each specific factive and non-factive verb has unique features of their meaning, ascriptions with each mental state verb did not always make sense in the context of the twelve different scenarios. Accordingly, for each scenario, we chose a pair of factive and non-factive terms that each made clear conceptual sense in the context described. Each factive and non-factive term was used in three scenarios.

Attitude	Factivity	Length	Frequency
'realize'	factive	7	.00338
'recognize'	factive	9	.00395
'understand'	factive	9	.01518

Attitude	Factivity	Length	Frequency
'is aware'	factive	8	.00610
'believe'	non-factive	7	.01716
'guess'	non-factive	5	.00339
'imagine'	non-factive	7	.00393
'assume'	non-factive	6	.00414

Table 1. Length and Frequency for factive and non-factive terms used in Experiment 4

5.3 Results

As in the previous experiments, data were excluded at the participant- and trial-level, and then analyzed using an identical set of linear mixed-effects models, only now the effects calculated for Ascription Type reflect whether a verb is factive vs non-factive rather than the verb being 'know' vs. 'think'[5].

This revealed a small main effect of Information Condition, $\chi^2(2) = 7.143$, $p = 0.028$, and no Information Condition × Ascription Type interaction, $\chi^2(2) = 0.686$, $p = 0.709$. More importantly, we observed a highly significant main effect of Ascription Type, $\chi^2(1) = 6.39$, $p < 0.011$, such that participants were faster to correctly assess the truth of ascriptions involving factive attitudes ($M = 2363.93$ms, $SD = 706.93$) than ascriptions involving non-factive attitudes ($M = 2452.7$ms, $SD = 736.3$) (see Figure 6).

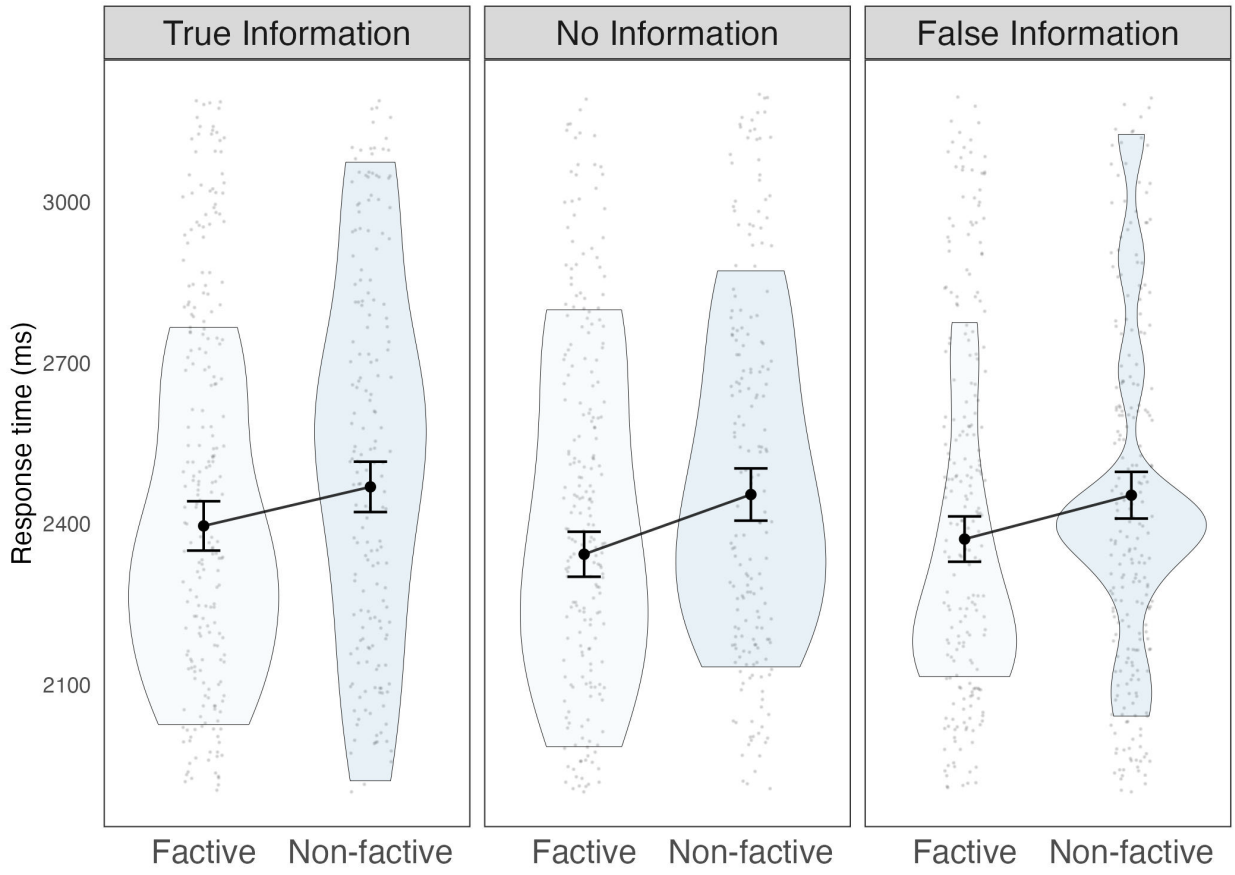

Figure 6. Response times for correct evaluations of factive mental state ascriptions (left points and violins)

[5]The fixed and random effects structure for the full model was specified as: response.time ∼ info.condition ∗ factivity.type + (1|scenario) + (info.condition ∗ factivity.type|subj). We were not able to include random slopes for the scenario because the crossed nature of the random effects in our experiment prevented convergence.

and non-factive mental state ascriptions (right points and violins) as a function of Information Condition (separate panels). Small points indicate trial-level responses, violins illustrate scenario- or item-level distributions, large dark points depict overall means, and error bars depict +/- 1 SEM.

5.4 Discussion

These results provide evidence that the response time difference observed for evaluations of knowledge and belief may generalize to the broader classes of factive and non-factive mental state verbs to which they respectively belong. This finding supports a more general pattern according to which factive mental state representations may be simpler because the content represented is necessarily consistent with one's own understanding (Phillips & Norby, 2019). Accordingly, these data also point toward a potential explanation of why evaluations of knowledge occur more quickly than evaluations of belief. However, regardless of whether or not this turns out to be the correct explanation for why evaluations of knowledge are faster than evaluations of belief, the previous four experiments collectively provide ample evidence that people do in fact make evaluations of knowledge without yet having made similar evaluations of belief.

6 Experiment 5

In the previous four studies, we investigated the relationship between knowledge and belief by examining the speed with which people were able to evaluate claims about others' mental states. We argued that the fact that people can correctly decide what others know faster than they can correctly decide what others think, provides evidence that they are not consulting their judgments of belief when forming their judgments about knowledge. Here, we ask a similar question but use a very different methodology. Specifically, we use functional magnetic resonance imaging (fMRI) to examine the neural responses exhibited during the formation of knowledge and belief representations and ask whether these neural patterns provide convergent evidence about the relationship between knowledge and belief.

To do this, we took advantage of an existing dataset from an experiment that examined the role of mental state representations in moral judgments for both neurotypical participants and participants with Autism Spectrum Disorder (ASD) (Chakroff et al., 2016). In the original study, participants read short stories involving different kinds of moral violations while undergoing functional magnetic resonance imaging. Critically for our purposes, these short stories manipulated the description of the agent's mental states. In some cases, the agent was described as having knowledge of some fact in the scenario (e.g., "knew the brakes were still broken"); in other cases, the agent was instead described as having a belief (e.g., "thought the detector just needed new batteries"); and in other cases, the agent was instead described as having some other factive mental state (e.g., "heard that he was in a relationship"; "realized the pond was unsafe").

While the original research focused on the neural processes underlying moral judgment, we can take advantage of the design used in this study by comparing the patterns of neural activation when participants learned about the agents' knowledge, belief, and other factive attitudes. Following the literature, we will focus on the neural response within the right temporo-parietal junction (RTPJ), which is well-known to play a highly selective and critical role in the neural computations underlying the representations of others' mental states (Gobbini et al., 2007; Koster-Hale & Saxe, 2013; Saxe et al., 2004). This area has even been shown, for example, to carry information about the specific kind of mental state being represented; allowing for one to decode whether the agent is represented as having auditory or visual evidence about some fact, among other things (Koster-Hale et al., 2014; Zaitchik et al., 2010). Our novel reanalysis of these data allows us to test two opposing predictions about the neural activity that knowledge and belief representations will elicit in RTPJ. On the one hand, if representing others' knowledge requires representing their beliefs, then we would expect the activity within RTPJ to reflect this structure. That is, while knowledge representations may require a number of different or additional neural processes, representations of knowledge should certainly not elicit less of a response in RTPJ than representations of belief, since they are meant to recruit precisely those representations.

On the other hand, the results of the previous five experiments suggest that representations of knowledge may not require representations of belief, and in fact may involve simpler or more basic processes (Phillips et al., 2020). Previous work has demonstrated that complex or effortful inferences typically elicit greater neural activity in RTPJ than simpler or less effortful inferences (Cohen et al., 1997; Meyer et al., 2015). In line with this, an alternative prediction is that representations of knowledge may elicit a similar amount, or even less, neural activity in RTPJ.

6.1 Participants

Participants included 26 neurotypical adults (NT; M_{age}=28.73; SD_{age}=12.01; 7 females) and 17 adults with high-functioning autism or Asperger's syndrome (ASD; M_{age}=31.41; SD_{age}=7.8; 2 females) recruited from the Greater Boston Area. All ASD participants received their diagnosis based on the Autism Diagnostic Observation Schedule, Second Edition (criterion $\geq$ 7; M = 9.60), as well as an impression by a trained clinician based on the diagnostic criteria of the DSM-IV.

6.2 Stimuli and procedure

As described in Chakroff et al. (Chakroff et al., 2016), participants read 60 stories depicting moral violations while undergoing fMRI. The stories were presented in cumulative segments. The last segment depicted the agent's mental state and was presented for 4 seconds. The agent's mental state was described using three different verb categories: knowledge, belief, or other factive attitude verbs. Word count was matched across conditions. Stories were presented in a pseudorandom order, divided into six 5.5 minute runs. To identify the RTPJ, all participants also completed a theory of mind functional localizer task (Dodell-Feder et al., 2011). This task consists of 10 stories about mental states (false-belief condition) and 10 stories about physical representations (false-photograph condition). (See http://saxelab.mit.edu/superloc.php for the task files). The task was presented in two 4.5 min runs, interleaved with the main experimental runs. Complete details of stimuli and procedure can be found in Chakroff et al. (Chakroff et al., 2016).

6.3 Results

As in Dodell-Feder et al. (2011), a whole-brain random effects analysis contrasting neural response in false-belief vs. false-photograph conditions ($p < 0.001$, uncorrected, $k > 16$) revealed activity in the RTPJ (peak voxel MNI coordinates: x = 58, y = -50, z = 28). We identified the RTPJ in all 38 participants individually, defined as contiguous voxels within a $9mm$ radius of the peak voxel that passed the contrast threshold.

We averaged the blood oxygen level dependent (BOLD) response across voxels in RTPJ from each subject during the 4 second segment in which the agent's mental state information was presented, and asked if this varied as a function of verb category. For each verb category, we calculated the average percent signal change (PSC) from baseline in RTPJ [PSC = 100 $\times$ raw BOLD response for (condition $-$ baseline) / raw BOLD response for baseline]. Baseline was defined as the average RTPJ response across all fixation time points between stimuli, adjusted for hemodynamic lag. Complete details of how the fMRI data were processed and analyzed can be found in Chakroff et al. (2016).

We then performed a linear mixed effects analysis predicting PSC in BOLD response in RTPJ, including both verb category (think verbs vs. knowledge verbs vs. other factive verbs) and participant group (autism vs. control) and their interaction.[6] We observed no verb category $\times$ group interaction ($\chi^2(2) = 0.31$, $p = 0.856$), indicating that neurotypical participants and participants with autism demonstrated similar activity across verb categories. More importantly, the analysis also revealed a main effect of verb category on RTPJ activity ($\chi^2(2) = 9.36$, $p = 0.009$) (Figure 7).

[6]Because of the fully crossed design used in Dodell-Feder et al. (2011), we were able to use maximal random effects structure for this analysis, i.e., RTPJ.BOLD $\sim$ mental.state $*$ ASD.group + (mental.state|Subject) + (mental.state|Item), where item codes the complement of the of the attitude verb used.

Pairwise comparisons of estimated marginal means revealed higher PSC for 'think' (EMM = 0.382, 95%CI: [0.282,0.482]), than for either 'know' (EMM = 0.245, 95%CI: [0.133,0.358]), t.ratio(19.93) = -2.62, p = 0.042, or other factive verbs (EMM = 0.266, 95%CI: [0.175,0.356]), t.ratio(17.75) = -2.68, p = 0.039. We did not find a significant difference between knowledge and other factive attitudes, t.ratio(25.54) = -0.46, p = 0.891.

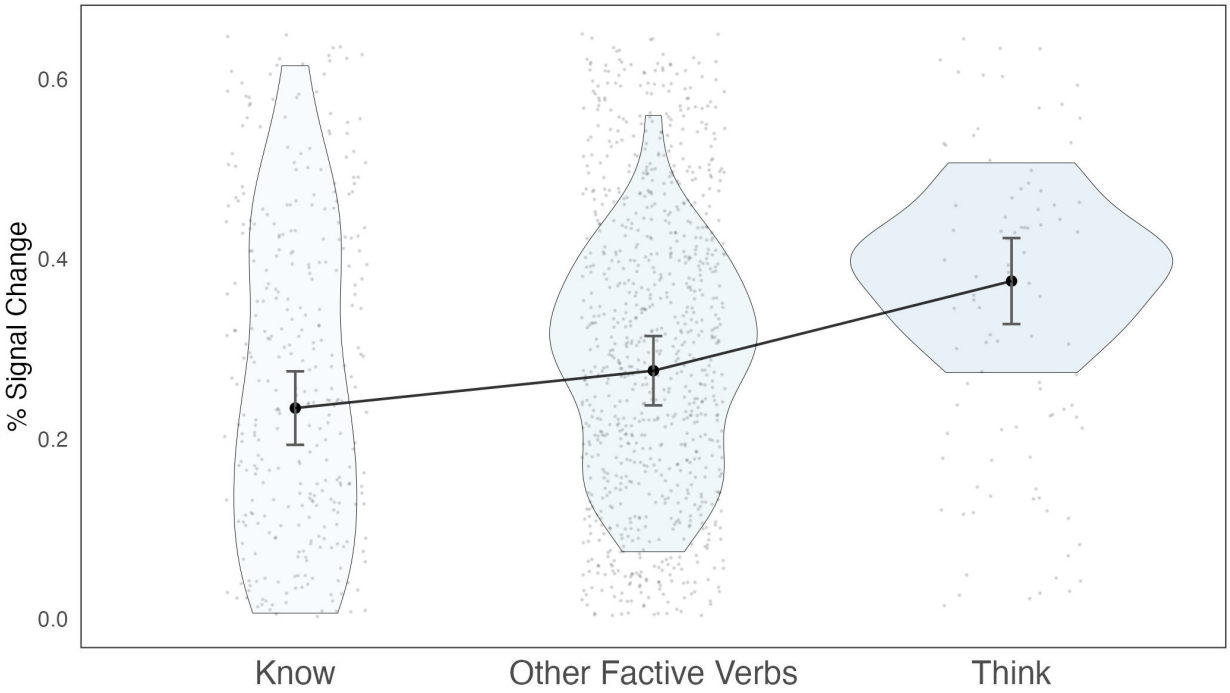

Figure 7. Percent signal change in RTPJ for each mental state verb: 'know' (left points and violin), other factive verbs, e.g., 'saw' or 'realized' (middle points and violin) and 'think' (right points and violin). Small points indicate trial-level responses, violins illustrate scenario- or item-level distributions, large dark points depict overall means, and error bars depict +/- 1 SEM.

6.4 Discussion

Our reanalysis of the fMRI data from Chakroff and colleagues (Chakroff et al., 2016) revealed that RTPJ exhibited a higher level of activity when participants read about others' beliefs than when participants read about others' knowledge (or factive attitudes more generally). This pattern clearly suggests that when participants were representing agents as knowing some proposition, they were not recruiting the set of processes that they employed when representing agents' beliefs. If they had been, then we should have observed at least a roughly similar level of activation in RTPJ. This result is also broadly in line with the findings by Zaitchik and colleagues (Zaitchik et al., 2010), who found greater activation across the theory of mind network when participants read sentences involving ascriptions of representational mental states (about an agent 'believing' or 'thinking' something) than when they read sentences involving perceptual or emotional states (e.g., that the agent 'saw' something, or what the agent was 'furious' about something). Critically, for our purposes, the perceptual and emotion mental states used were largely factive, while the representational mental states used were largely non-factive (Zaitchik et al., 2010).[7] Our finding of lower BOLD responses in the RTPJ when representing an agent as knowing as compared to believing is also consistent with the EEG results of Bricker (2020), who found reduced inhibition of the P3b amplitude when participants attributed beliefs to another agent as compared to knowledge. Finally, in line with the findings in our Experiment 2, we also do not find differences in the overall pattern observed between neurotypical participants and participants with ASD.

[7]Zaitchik and colleagues lumped the term 'know' with the representational mental states and did not directly compare knowledge ascriptions to other kinds of representational mental states, as we do here.

This pattern of results strengthens the case for the conclusions we've drawn across the previous studies. Despite the highly different methodologies used, both approaches yielded consistent evidence, which showed that knowledge representations do not rely on belief representations and that knowledge representations may actually be simpler than belief representations. Moreover, the differences between these two approaches complement each other in another way: while alternative proposals may be offered to explain either one of these sets of results, it is very unlikely that any proposal aside from ours can easily explain and predict the pattern of results we observed across both studies (i.e., that representations of knowledge are computed more quickly and require less neural activity in classic theory of mind brain regions).

7 General Discussion

Across five experiments, we found consistent evidence that the representation of others' knowledge can occur in the absence of representations of corresponding beliefs. Experiment 1 demonstrated that accurate evaluations of knowledge ascriptions happen significantly faster than accurate evaluations of otherwise identical belief ascriptions. Experiment 2 further found that this effect cannot be explained by pragmatic differences and that it extends to participants with Autism Experiment 3 demonstrated this finding cross-linguistically, establishing that people also accurately evaluate others' knowledge faster than their beliefs in French. Experiment 4 provided evidence that this response time difference generalizes to the larger classes of factive and non-factive attitudes to which knowledge and belief respectively belong. Finally, Experiment 5 revealed that forming representations of an agents' knowledge (or other factive mental states) elicits less activity in classic theory of mind brain regions than forming representations of that agent's beliefs. Taken together, these studies provide evidence for the *knowledge without belief* hypothesis. That is, they provide evidence that there are cases in which people form representations of another person's knowledge without forming representations of that person's belief.

7.1 Relation to previous research

Previous research has provided evidence that the capacity to represent knowledge may be present before the capacity to represent belief (see (Nagel, 2017; Phillips et al., 2020) for reviews and discussion). However, this prior work was largely indirect. What it directly showed was that there are certain kinds of cases in which the early capacity to understand others' minds seems to be limited, and those cases happen to align with when we would naturally say that someone "believes" something rather than "knows" something. For example, one finds the capacity to represent others as mistakenly believing something to be the case to emerge later than the capacity to represent someone as understanding *true* things about the world. This characteristic limit appears to emerge in infants (Luo & Johnson, 2009; Vouloumanos et al., 2014), in automatic processes within adults (A. Surtees et al., 2016; A. D. Surtees & Apperly, 2012), and in monkeys (Horschler et al., 2019, 2020; Martin & Santos, 2014, 2016).

One natural way to explain this broad pattern is to argue that the capacity to represent others as knowing something is present before the capacity to represent others as believing something. If this hypothesis were right, then one should obviously expect earlier success on theory of mind tasks when the agent's mental state amounts to knowledge and later success on tasks where the agent agent does not actually know something, but merely believes it (Phillips et al., 2020).

At the same time though, there may be more deflationary explanations for why one finds earlier success on the tasks where the agent can be described as knowing rather than believing something. For example, it may simply be easier to represent someone's mental state when their understanding of the world does not conflict with your own (see, e.g., (Deschrijver & Palmer, 2020)). Perhaps then, what this prior work merely shows is that it is easier to represent others' true beliefs than it is to represent their false beliefs. This alternative explanation would have at least some potential to explain the results from previous research.

Critically, this kind of deflationary explanation does not apply to the studies we've presented, which differ in two critical ways from this prior work. First, our studies bear more directly on the relationship between

knowledge and belief in that we ask participants explicitly about what another agent "knows" or "thinks" (along with other related mental state terms). Second, we only ask participants to evaluate whether the agent knows or thinks things that the participants themselves take to be true. In some conditions, such as the True Information condition, the agent both thinks and and knows these things, while in others (No/False Information conditions), the agent neither thinks nor knows these things. Across these cases we still find that participants are faster to correctly evaluate whether another agent knows that thing than they are to correctly evaluate whether another agent believes that thing. These results thus provide more direct evidence for the knowledge without belief hypothesis.

7.2 Interactions between Ascription Type and Information Condition

Across a number of studies (Experiments 1-3), we also found an interaction effect such that the greatest difference in response time in evaluating knowledge and belief ascriptions was in the No Information condition, while a somewhat smaller difference was observed in the True Information condition, and the smallest difference was consistently observed in the False Information condition. While any proposed explanation of this pattern will necessarily be speculative until further research is done, offering some way of understanding this pattern may still be desirable, given that we observe it relatively consistently. Thus far, we've argued that correct knowledge evaluations may be faster than correct belief evaluations because evaluations of knowledge may involve simpler or more basic processes. What we now need to explain is why this response time difference is moderated in some cases. One potentially helpful place to start is to recall that knowledge typically entails belief, and so in any case where you recognize someone as knowing something, you should be able to use that fact to infer that they must also believe that thing (Ichikawa & Steup, 2016; Myers-Schulz & Schwitzgebel, 2013; Rose & Schaffer, 2013). Accordingly, in the True Information condition, where the agent does in fact have knowledge, participants should be able to relatively easily infer that the agent also has the corresponding belief. In contrast, in the No Information condition, when the agent does not have knowledge, participants cannot use this strategy, and they have to independently evaluate whether the agent has the relevant belief. This explanation makes the more precise prediction that the smaller difference in the True Information condition occurs in part because participants are faster to correctly evaluate what the agent thinks in the True Information condition than in the No Information condition. Our data are consistent with this prediction: combining the data from Experiments 1-3, we found that correct evaluations of belief ascriptions were made faster in the True Information than in the No Information condition, $t(1032) = -3.83$, $p < .001$.

What remains to be explained is why we observe a yet smaller difference in the False Information condition. Importantly, the vignettes used in these conditions explicitly required participants to represent the agent as believing some incorrect proposition, *q*. Recall, however, that we asked participants to evaluate whether or not the agent thinks or knows not *q*, but instead a true proposition *p* (ensuring that the correct answer was the same for both types of ascriptions, as they neither believe, nor do they know, *p*). Thus, the background vignettes used in these conditions require participants to engage in reasoning about what the agent (falsely) believes. Having previously represented the agent as believing *q*, participants may find it particularly easy to infer that the agent does not believe *p* (as *p* and *q* are always inconsistent). Separately, in this case, determining that the agent does not know *p* may be particularly effortful in this case, as *p* is true.

While we think this series of explanations is a plausible way of explaining the observed interaction effects, future work should further pursue the potential interaction between the information state of an agent and representations of knowledge and belief.

7.3 Conclusion

This paper set out to ask whether people can represent someone else as knowing something without representing them as believing that thing. Across a series of five studies using a variety of manipulations and methods, we found strong evidence for an affirmative answer to this question: representations of what others do or do not know can occur in the absence of representations of what they believe. This series of findings

provides further support for the growing consensus that the capacity for knowledge representation is more basic than the capacity for belief representation.

References

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

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. Oxford university press.

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). lme4: Linear mixed-effects models using eigen and S4. *R Package Version*, *1*(7).

Bricker, A. M. (2020). The neural and cognitive mechanisms of knowledge attribution: An EEG study. *Cognition*, *203*, 104412.

Buckwalter, W., Rose, D., & Turri, J. (2015). Belief through thick and thin. *Noûs*, *49*(4), 748–775. https://doi.org/https://doi.org/10.1111/nous.12048

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337–342.

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: Intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, *11*(3), 476–484.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*(6625), 604–608.

Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting*, *32*(1), 119–134.

Deschrijver, E., & Palmer, C. (2020). Reframing social cognition: Relational versus representational mentalizing. *Psychological Bulletin*.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *Neuroimage*, *55*(2), 705–712.

Dudley, R. (2018). Young children's conceptions of knowledge. *Philosophy Compass*, *13*(6), e12494.

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-olds' understanding of know and think. In *Experimental perspectives on presuppositions* (pp. 241–262). Springer.

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, *19*(11), 1803–1814.

Harris, P. L., Yang, B., & Cui, Y. (2017). "I don't know": Children's early talk about knowledge. *Mind & Language*, *32*(3), 283–307.

Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi, K., Kato, S., Hori, Y., Nagai, Y., Iijima, A., Someya, T., & Hasegawa, I. (2020). Macaques exhibit implicit gaze bias anticipating others' false-belief-driven actions via medial prefrontal cortex. *Cell Reports*, *30*(13), 4433–4444.e5. https://doi.org/https://doi.org/10.1016/j.celrep.2020.03.013

Heim, I. (1991). Artikel und definitheit. *Semantik: Ein I Nternationales Handbuch Der Zeitgenössischen Forschung*, 487–535.

Hirschberg, J. L. B. (1985). *A theory of scalar implicature*. University of Pennsylvania.

Holland, C., & Phillips, J. (2020). A theoretically driven meta-analysis of implicit theory of mind studies: The role of factivity. *CogSci*.

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

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2019). Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. *Cognition*, *190*, 72–80.

Ichikawa, J. J., & Steup, M. (2016). The analysis of knowledge. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University.

Kissine, M. (2012). Pragmatics, cognitive flexibility and autism spectrum disorders. *Mind & Language*, *27*(1), 1–28. https://doi.org/https://doi.org/10.1111/j.1468-0017.2011.01433.x

Koster-Hale, J., Bedny, M., & Saxe, R. (2014). Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition*, *133*(1), 65–78.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron*, *79*(5), 836–848.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834.

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.

Lakoff, R. (1973). *Language*, *49*(3), 685–697. https://doi.org/https://doi.org/10.2307/412359

Lenth, R. V. (2023). *Emmeans: Estimated marginal means, aka least-squares means.* https://CRAN.R-project.org/package=emmeans

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind." *Trends in Cognitive Sciences*, *8*(12), 528–533.

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

Martin, A., & Santos, L. R. (2014). The origins of belief representation: Monkeys fail to automatically represent others' beliefs. *Cognition*, *130*(3), 300–308.

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

Meyer, M. L., Taylor, S. E., & Lieberman, M. D. (2015). Social working memory and its distinctive link to social cognitive ability: An fMRI study. *Social Cognitive and Affective Neuroscience*, *10*(10), 1338–1347.

Myers-Schulz, B., & Schwitzgebel, E. (2013). Knowing that p without believing that p. *Noûs*, *47*(2), 371–384.

Nagel, J. (2017). Factive and nonfactive mental state attribution. *Mind & Language*, *32*(5), 525–544.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., & Knobe, J. (2020). Knowledge before belief. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20000618

Phillips, J., & Norby, A. (2019). Factive theory of mind. *Mind & Language*. https://doi.org/10.1111/mila.12267

Rose, D., & Schaffer, J. (2013). Knowledge entails dispositional belief. *Philosophical Studies*, *166*(1), 19–50.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.*, *55*, 87–124.

Stich, S. (2013). Do different groups have different epistemic intuitions? A reply to jennifer nagel 1. *Philosophy and Phenomenological Research*, *87*(1), 151–178.

Surtees, A. D., & Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. *Child Development*, *83*(2), 452–460.

Surtees, A., Apperly, I., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition*, *150*, 43–52.

Svindt, V., & Surányi, B. (2021). The comprehension of grammaticalized implicit meanings in SPCD and ASD children: A comparative study. *International Journal of Language & Communication Disorders*, *56*(6), 1147–1164.

Vouloumanos, A., Martin, A., & Onishi, K. H. (2014). Do 6-month-olds understand that speech can communicate? *Developmental Science*, *17*(6), 872–879.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Williamson, T. (2002). *Knowledge and its limits.* Oxford University Press.

Yuan, Y., & Kim, M. (2021). Cross-cultural convergence of knowledge attribution in east asia and the US. *Review of Philosophy and Psychology*, 1–28.

Zaitchik, D., Walker, C., Miller, S., LaViolette, P., Feczko, E., & Dickerson, B. C. (2010). Mental state attribution and the temporoparietal junction: An fMRI study comparing belief, emotion, and perception. *Neuropsychologia*, *48*(9), 2528–2536.