# The Neural and Cognitive Mechanisms of Knowledge Attribution: An EEG Study

Adam Michael Bricker

A Thesis for the Master's Degree Program in Human Neuroscience
Supervision: Henry Railo
Faculty of Medicine, University of Turku
Spring 2020

## Abstract

Despite the ubiquity of knowledge attribution in human social cognition, its associated neural and cognitive mechanisms are poorly documented. A wealth of converging evidence in cognitive neuroscience has identified independent perspective-taking and inhibitory processes for belief attribution, but the extent to which these processes are shared by knowledge attribution isn't presently understood. Here, we present the findings of an EEG study designed to directly address this shortcoming. These findings suggest that belief attribution is not a component process in knowledge attribution, contra a standard attitude taken by philosophers. Instead, observed differences in P3b amplitude indicate that knowledge attribution doesn't recruit the strong self-perspective inhibition characteristic of belief attribution. However, both belief and knowledge attribution were observed to display a late slow wave widely associated with mental state attribution, indicating that knowledge also shares in more general processing of others' mental states. These results provide a new perspective both on how we think about knowledge attribution, as well Theory of Mind processes generally.

Keywords: Theory of Mind, Mentalizing, Perspective Taking, Self-Perspective Inhibition, Knowledge-First, Epistemology

## 1. Introduction

From as early as we begin to use mental state terms, knowledge attribution is a staple of human social cognition (see e.g. Shatz et al. 1983). Throughout the course of daily life, we frequently attribute knowledge to others, and experimental evidence indicates that we then base our own beliefs and actions on what we take others to know (e.g. Turri 2015; Turri 2017; Turri et al. 2017). The human significance of knowledge attribution is reflected in the substantial philosophical effort devoted to describing when we do and don't attribute knowledge to others, the origins of which date back millennia to Plato's *Meno* and *Theaetetus* (see Ichikawa and Steup 2018). Despite this, the mechanisms underlying knowledge attribution have attracted remarkably little attention and are not presently understood. In the past two decades, cognitive neuroscience has made significant progress in understanding the neural and cognitive processes underlying the human capacity to accurately evaluate and reason about the mental states of others, referred to collectively as "Theory of Mind" (ToM; for reviews, see Carrington & Bailey 2009; Mahy et al. 2014; Heleven & Van Overwalle 2018). However, this research typically uses belief as its paradigm mental state, and knowledge attribution has gone systematically overlooked. The extent to which knowledge attribution relies on the same neural and cognitive mechanisms as belief attribution is presently unclear. It is the aim of the present study to address this shortcoming.

### *1.1 Knowledge and Belief States*

At the conceptual level, this study will employ a broad characterization of knowledge and belief states, which derives from widely recognized characteristics of knowledge within epistemology (the philosophy of knowledge). Note that this characterization will understand knowledge in a strictly propositional sense (i.e. knowledge that something is the case).

When we judge that someone is in a knowledge state, we attribute to them, roughly speaking, a representation of the state of reality that matches how reality actually is (that is to say, is true). This reflects what is often referred to as the "factivity" of knowledge, and it is all-but universally considered to be a defining characteristic of knowledge states. As put by Turri, "One of the very few things that epistemologists of all stripes agree on is that knowledge requires truth" (2011, 141). We might contrast this requirement with attributions of belief states. When we say that someone is in a belief state, while we also attribute a state that in some sense represents reality, this is a state that may or may not match the state of reality (i.e. could be true or false).

Crucially, however, when we judge that someone is in a knowledge state, we don't just attribute to them a representation of the state of reality that matches reality—we attribute a representation that, in some sense, *stably* matches reality. While the precise nature of this stability is not widely agreed upon, the general idea proliferates epistemological thought, and might be illustrated by appeal to a few key examples: For Unger, this meant that "it is not at all accidental" that the state matches reality (1968, 158). Goldman expressed this stability in terms of knowledge not just matching reality in one's actual situation, but also in "relevant counterfactual situations" (1976, 771). Pritchard, effectively merging these ideas, frames non-accidentally matching reality in terms of matching reality in relevant counterfactuals (2005). While the epistemology here is much more complex than indicated, the finer details aren't especially relevant for the present study. The important point is simply that, when we attribute a knowledge state to someone, we attribute a state that doesn't merely reflect reality. We attribute to them a state that displays some stability in its reflection of reality.

While this characterization allows us to understand something about *what* we attribute when we attribute knowledge, it remains unclear *how* we attribute knowledge. In order to address this question, as with the *what* of knowledge, the present study will also seek to characterize the *how* in relation to the *how* of belief attribution. This approach will be facilitated by the large body of empirical research on the neurocognitive mechanisms of belief attribution, which will form the theoretical basis for this study.

### *1.2 Neural and Cognitive Mechanisms of Belief Attribution*

Converging hemodynamic and neuropsychological evidence indicates that there are two distinct neurocognitive processes that follow perspective computation during belief attribution: (i) self-perspective inhibition and (ii) other perspective taking.

A number of lesion studies have indicated a double dissociation between perspective taking and self-perspective inhibition (Apperly et al. 2004; Samson et al. 2004, 2005, 2007). For example, a pair of studies from Samson et al. explored when patients with either left temporoparietal junction (TPJ) or frontal lesions could correctly infer an agent's belief about an object's location, both when patients could and couldn't see the objects location (2004; 2005). While patients with damage to the TPJ were unable to accurately to attribute belief of an object's location, this inability both extended to trials where patients lacked independent information about the object's location and did not appear to be explained by any other deficits in inhibitory control (Samson et al. 2004). Conversely, a patient with right prefrontal and temporal damage displayed poor performance in tasks where the agent's belief was inconsistent with the patient's perspective information but was successful in attributing belief when independent perspective information was absent (Samson et al. 2005). On this basis, Samson et al. concluded that "the inhibition of one's own point of view and the ability to infer someone else's point of view rely on distinct neural and functional processes" (2005, 1102).

These neuropsychological findings have been confirmed by a wealth of observations via functional imaging, which have identified distinct hemodynamic correlates for other perspective taking and self-perspective inhibition during belief attribution. A number of fMRI studies have linked the TPJ with perspective taking specifically during belief attribution (Schurz et al. 2013; Schuwerk et al. 2014; Schurz et al. 2015; Özdem et al. 2019; see also Heleven & Van Overwalle 2018). For example, a study by Schuwerk et al. used a two-phase design to isolate the process of perspective taking from antecedent belief computation (2013). In the first phase, participants watch a short video story in which an agent formed either a true or false belief about an object's location; in the second phase, participants were asked to report on either their own belief or the agent's belief about the object's location. While the TPJ was recruited for both the computation and the perspective-taking phases, only the TPJ showed increased activation during perspective taking. Conversely, multiple fMRI studies have also reported that the inferior frontal gyrus (IFG) plays a specialized role in self-perspective inhibition during belief attribution (van Der Meer et al. 2011; Hartwright et al. 2012; Hartwright et al. 2015). The consistency of these neural correlates with the regions identified from the neuropsychological studies strengthens the hypothesis that perspective taking and self-perspective inhibition are distinct neurocognitive processes.

Additional evidence from functional imaging and neuropsychology indicates that the inhibitory control recruited during mental state attribution is a unique neurocognitive process that dissociates from more generalized inhibition. Using a design in which participants made both mental (i.e. belief attribution) and non-mental judgements under conditions requiring either high or low inhibition of self-perspective information, an fMRI study by Hartwright et al. found that the ventrolateral prefrontal cortex (vlPFC) displayed greater activation between high-inhibition and low-inhibition conditions, but only during belief attribution (2015). Moreover, a neuropsychological study from Samson et al. observed a double dissociation between self-perspective inhibition during desire attribution and non-mental inhibition: Patients with deficits in self-perspective inhibition didn't necessarily display decreased inhibitory control regarding non-self-perspective information, and patients with general deficits in inhibitory control didn't necessarily display deficits in self-perspective inhibition (2015).

Taken together, these findings constitute strong evidence that perspective taking and self-perspective inhibition during belief attribution are two distinct neurocognitive processes, with distinct hemodynamic correlates. However, the extent to which *knowledge* attribution might share either or both of these processes is not presently clear.

The electrophysiological correlates of belief attribution are not as well understood. A large number of EEG studies have observed late slow waves (LSWs) associated with mental state attribution, which characteristically begin at least 500 ms after stimulus onset, and can display both frontal and parietal scalp distributions (Sabbagh & Taylor 2000; Liu et al. 2004; Liu et al. 2009a/b; Zhang et al. 2009; McCleery et al. 2011; Meinhardt et al. 2011; Chen et al. 2012; Geangu et al. 2013). Earlier differences between ToM and non-ToM have been also observed for the frontal LSW (Sabbagh & Taylor 2000). However, when earlier differences have been observed parietally that are consistent with the P3 component (e.g. Sabbagh et al. 2004; Meinhardt et al. 2011), these have been previously characterized as differences in the P3 component, not the LSW. Accordingly, here LSWs will be individuated from other effects both by latency (> 500ms) and spatial distribution (frontal + parietal).

Despite the large number of studies reporting LSWs in association with mental state attribution, it is not clear if this LSW component is correlated with self-perspective inhibition, perspective taking, or both. Zhang et al. have concluded that the LSW corresponds with false belief reasoning, maintaining that self-perspective inhibition occurs prior to LSW onset (2009). Conversely, McCleery et al. have concluded that the LSW is associated with inhibitory control during belief attribution (2011). However, both these conclusions rest on questionable bases. Zhang et al. base their conclusions largely on the differences in P3 amplitudes between reasoning about true and false beliefs, together with the fact that the P3 component has been proposed to reflect an inhibitory mechanism (2009). We might question the strength this conclusion given both the distinct neurocognitive status of self-perspective inhibition (discussed above) and evidence that the P3 specifically involves inhibition of *extraneous* neural processing (see Polich 2011; discussed further in §1.4). Similarly, while McCleery et al. base their conclusion on the frontal distribution of the LSW, their own source reconstruction shows IFG activation only in "consistent" conditions, in which demand for self-perspective inhibition should be lower (2011). By exploring the neurophysiological correlates of previously unexamined attributions, like knowledge, one of the aims of this study is to provide new evidence regarding whether the LSW component is associated with self-perspective inhibition and/or perspective taking.

### *1.3 Modeling Knowledge Attribution*

In order to understand how self-perspective inhibition and perspective taking might play a role in knowledge attribution, we need some way of modeling how belief attribution relates to knowledge attribution. There is precedent for two preliminary approaches to modeling the cognitive stages of attributing knowledge to others: (i) a composite approach and (ii) a mental state approach. These approaches roughly track a key distinction between how philosophers and psychologists think about knowledge: Many philosophers of knowledge (i.e. "epistemologists") don't tend to count knowledge among paradigm mental states like belief and desire, whereas psychologists often do.

On one approach often favored in epistemology, knowledge attribution is modeled as a complex judgement consisting of both belief attribution and additional judgements about non-mental states. At a minimum, these judgements about non-mental states will include a judgement about the state of reality: Is the belief true? Additionally, the non-mental component will also include judgements reflecting whatever separates true belief from knowledge. Let's call this picture of knowledge attribution the "composite model" (CM). Importantly, on this model, belief attribution is conceptualized as a stage of knowledge attribution: In the course of attributing knowledge to another person, we first evaluate whether they are in the mental state of believing, and then make a series of non-mental judgements about the truth and provenance of that belief. In modeling belief attribution as a stage of knowledge attribution, we can understand that the CM entails that both self-perspective inhibition and perspective taking are components of knowledge attribution. The CM is the model of knowledge attribution suggested by the standard way philosophers think about the relationship between belief and knowledge. Many epistemologists do not think that knowledge is a mental state, and Nagel has even so far as to contend that "the current climate in philosophy is largely hostile to the idea that knowledge is a mental state" (2013, 274). Instead, they take knowledge to consist of a belief state taken in conjunction with additional non-mental conditions. The CM reflects this epistemological perspective by modeling belief attribution as a stage of knowledge attribution.

Conversely, the standard view of psychologists is that knowledge is a mental state in its own right: In attributing knowledge to others, we use our ToM systems to directly evaluate whether they are in the mental state of knowing. Importantly, on this mental state approach, belief attribution is not conceptualized as a stage of knowledge attribution, but rather as an analogous process. This approach to knowledge attribution can be found in the tendency of empirical literature to often list knowledge alongside belief and desire as a paradigm mental state (see e.g. McCleery et al. 2011; Bradford et al. 2015; Ferguson et al. 2017; Hyde et al. 2018). Additionally, two prominent philosophers have also argued that knowledge is a mental state (Williamson 2000; Nagel 2013). Despite this, however, philosophical support for the view that knowledge is a mental state remains limited (see Fricker 2009; Smith 2017).

Following the mental state approach, there are at least two straightforward ways we might model knowledge attribution. The first we might call the "belief-like model" (BlM), on which the neurocognitive processes that guide belief attribution are shared by knowledge attribution. Importantly, on the BlM, both self-perspective inhibition and perspective taking are component processes of knowledge attribution. The BlM is the model implicit in the work of philosophers like Nagel (2010; 2013) and Gerken (2013), who have made broad theoretical claims about the cognitive architecture of knowledge attribution based primarily on empirical findings from belief attribution. In attempting to explain certain patterns of knowledge attribution, Nagel has even maintained that self-perspective inhibition is required for accurate knowledge attribution, primarily on the basis of the role such inhibitory control plays in the accuracy of belief attribution (2010).

Additionally, still taking the mental state approach, there is a second way we might model knowledge attribution. Knowledge attribution differs from belief attribution in the crucial respect that self-perspective information can be and usually is relevant in determining whether others have knowledge, even when that information is inconsistent with other-perspective information. Again, in the minimal case, this self-perspective information will involve the truth about the state of reality. Even (or perhaps especially) when the truth is not available from the other perspective, it is crucial that it is integrated into judgements about knowledge. On a speculative basis, one way this might be accomplished is by avoiding the same degree of self-perspective inhibition entailed by belief attribution. Instead, knowledge attribution might recruit less, or even no, self-perspective inhibition in order to avoid inhibiting epistemically relevant information. Let's call this the "weakly inhibitory model" (wIM) of knowledge attribution. On this model, unlike the CM and BlM, knowledge attribution is characterized by weak-to-no self-perspective inhibition during perspective taking.

| | *Belief attribution stage?* | *Strong self-perspective inhibition?* |
|---|---|---|
| *Composite Model (CM)* | Yes | Yes |
| *Belief-like Model (BlM)* | No | Yes |
| *Weakly Inhibitory Model (wIM)* | No | No |

*Table 1: Summary of key differences between the three preliminary models of knowledge attribution. Note that on all three, knowledge attribution is modeled as recruiting the perspective taking characteristic of attributing mental states to others.*

### *1.4 CM, BlM, and wIM: Behavioral and Electrophysiological Predictions*

The behavioral predictions of the CM diverge from those of the BlM and wIM in at least one crucial respect. On the CM, belief attribution is modeled as one stage of knowledge attribution. Thus, the CM predicts longer reaction times (RTs) for knowledge attribution than belief attribution for otherwise comparable ToM tasks. The BlM and wIM make no such predictions. Were we to observe equivalent RTs for belief and knowledge attribution, or shorter RTs for knowledge than belief, this would be inconsistent with the CM. Assuming that the composite and mental state approaches represent the only plausible means for modeling belief attribution, we might understand that this would then constitute evidence not only against the CM, but also in support of the mental state approach. Conversely, greater RTs for knowledge than belief would be consistent with the CM, BlM, and wIM.

The different cognitive demands entailed by knowledge attribution on the composite, belief-like, and weakly inhibitory models also correspond with diverging electrophysiological predictions, which might be tested via EEG. This study has been designed primarily to test those predictions associated with the amplitude of the P3 event-related potential (ERP) component. Roughly speaking, ERPs depict microvolt-scale changes in electrical potential related to some stimulus or task (i.e. "event"), which are generally measured at the scalp level. While these changes in potential can reveal millisecond-level changes in neural activity, the spatial resolution is not as precise. Moreover, as the signal-to-noise ratio for single-trial ERPs is quite low, they are usually averaged over tens or hundreds of trials in order to gain a usable signal. ERPs are especially useful for research into the brain and cognition, as they display a number of widely documented "components," deflections in potential with a characteristic polarity, latency, and scalp distribution, which index processing in the brain. Thus, by comparing differences in an ERP component between conditions (i.e. the component's amplitude or latency), we can infer differences in neural processing between conditions (for more, see Luck 2014).

As already indicated, the ERP component of primary interest to the present study is the P3. Ubiquitous for tasks requiring stimulus discrimination, the P3 component has been hypothesized to reflect a general inhibitory mechanism for "extraneous" neural activity during stimulus discrimination (Polich 2007, 2137). Note that while this inhibitory hypothesis isn't integral to the present account, it does provide a helpful explanatory framework for understanding the P3 component. There are a number of cognitive factors known to modulate P3 amplitude which the inhibitory hypothesis explains (see Polich 2011 for more): One such factor is target stimulus probability. P3 amplitude is notably greater for less frequent target stimuli (Duncan-Johnson & Donchin 1977), which we might understand as reflecting the fact that, as "low probability stimuli can be biologically important, it is adaptive to inhibit unrelated activity to promote processing efficiency" (Polich 2007, 2137).

Another classical P3 finding comes from dual-task paradigms: When participants are required to complete a discrimination task and some additional task simultaneously, the P3 amplitude for the discrimination task *decreases* with increasingly demanding additional tasks (Isreal et al. 1980; Kramer et al. 1983). On the inhibitory P3 hypothesis, this correlation can be understood in the following way: High processing demands for the additional task limit the available neural resources for the inhibitory mechanism responsible for the P3, thus resulting in a reduced P3 amplitude (Polich 2011). In addition to explicit dual-task studies, a similar P3 effect has been observed for other discrimination tasks that entail additional cognitive processes. The best example of this comes from n-back tasks, which, despite not strictly being "dual tasks," strongly recruit working memory concurrently with stimulus discrimination. As demands on working memory increase (i.e. as n increases), P3 amplitudes for the stimulus matching component of the n-back decrease (Watter et al. 2001; Covey et al. 2017; Pergher et al. 2019). As with explicit dual tasks, this can be understood as "reflecting a reallocation of attention and processing capacity away from the processes relative to which the P300 is generated" (Watter et al. 2001). This is the principle on which the electrophysiological predictions of this study are built: During stimulus discrimination, demand for neural resources from additional cognitive processes can reallocate resources away from the inhibitory mechanism that produces the P3, resulting in reduced P3 amplitudes.

The operative difference between, on the one hand, the composite and belief-like models and, on the other hand, the weakly inhibitory model is the extent to which self-perspective inhibition is recruited during knowledge attribution: On the wIM, knowledge attribution is characterized by an absence of the strong self-perspective inhibition entailed by the BlM and CM. This means that, when discriminating between visual stimuli (e.g. cartoons) on the basis of the knowledge states depicted by the stimuli, the BlM entails higher additional processing demands than the wIM, making the P3 component a plausible candidate for adjudicating between the BlM and wIM. Assuming that P3 amplitudes will correlate primarily with additional processing demands, similarly, e.g., to n-back tasks, the predictions of the CM and BlM diverge from those of the wIM in the following way: On both the CM and BlM, strong self-perspective inhibition engages concurrently with perspective taking during knowledge attribution, either as a component process of belief attribution (CM) or in a knowledge attribution process that recapitulates the neurocognitive mechanisms of belief attribution (BlM). In this context, "strong" will be understood in a functional, albeit circular sense, on which "strong self-perspective inhibition" entails a sufficiently high demand for neural resources to, at least when recruited in conjunction with perspective taking, systematically interfere with resource allocation for the P3 mechanism. Therefore, both the CM and BlM predict a reduced P3 amplitude for knowledge attribution, comparable to that of belief attribution. Conversely, the wIM doesn't entail any strong (self-perspective) inhibitory demands for knowledge attribution during perspective taking. On the weakly inhibitory model, if knowledge attribution recruits self-perspective inhibition at all, it does so to a significantly lesser degree than belief attribution, which wouldn't to the same degree interfere with resource allocation to the P3 mechanism. Therefore, the wIM predicts a higher P3 amplitude for knowledge attribution than for belief attribution. The behavioral and electrophysiological predictions of the CM, BlM, and wIM are summarized in table 1.

Finally, it should be noted that the present experiment has been carefully designed so that differences in P3 amplitude might be reasonably assumed to primarily reflect additional processing demands during stimulus discrimination. Most importantly, this means ensuring (i) tasks are kept simple enough to preclude P3 differences associated with discrimination difficultly or uncertainty, and (ii) target stimuli frequencies are kept similar enough between conditions to avoid P3 differences associated with differences in target probability (see Polich 2011; Luck 2014, ch. 3). The veracity of this assumption is itself empirically verifiable, as there are a number of observations that would be inconsistent with it (see §4.2). It should also be noted that this experiment assumes that in the case of belief attribution, the resource demand from self-perspective inhibition can indeed be great enough to reduce the resources available for the P3. While we have good theoretical reason to suppose this may be the case, previous EEG studies investigating belief attribution (see above) have generally avoided the straightforward stimulus discrimination design most conducive to such an effect. However, were it the case that belief attribution simply cannot produce this P3 effect, it would be immediately clear from the results of this study.

| | **P3: K > B** | **P3: K = B** |
|---|---|---|
| *RT: K > B* | wIM | CM or BlM |
| *RT: K ≤ B* | wIM | BlM |

*Table 2: Summary of Model Predictions for Knowledge (K) and Belief (B) Attribution*

## 2. Methods

### *2.1 Overview*

For each trial, participants viewed a simple illustration of a man (“Steve”) sitting at a table (see fig. 1). For any given trial, there were between zero to three cylinders illustrated on the table. Sometimes Steve couldn’t see all the cylinders from his perspective, but the entire table was always visible from the participant’s perspective. The experiment followed a simple 1 factor (judgement type) design with 4 levels: (i) judgement about non-mental state of reality; (ii) belief attribution; (iii) belief attribution + judgement about non-mental state of reality; (iv) knowledge attribution. Identical stimuli were used for every condition. The tasks were divided into blocks, with verbal instructions provided at the start of each block.

### *2.2 Participants*

Data from 32 participants were included in this study (8 male, 24 female; 4 left handed, 28 right handed). Mean participant age was 23.2 years (range: 19-32). Participants were all university students living in Turku, Finland. No compensation was offered for participation, but psychology students were entitled to received course credit. All participants self-reported that they had not previously been diagnosed with any neurological or psychiatric disorders. All participants provided full informed consent, in accordance with the Declaration of Helsinki.

The data of an additional 5 participants was excluded from this study, 2 due to a clear misunderstanding of at least one task (false-alarm rate > .75 each) and 3 due to poor recording quality.

### *2.3 Design and Procedure*

Before starting the experiment, participants were verbally provided with a background story to accompany the stimuli. Concurrently, specific stimuli were shown to the participant in order to illustrate certain points of this background story. The story conveyed the following information: Steve’s job is to sit at the table and count the number of cylinders. Sometimes Steve will be able to see the entire table, but other times he can only see part of the table. Steve's boss has told Steve that there will never be any cylinders that he cannot see. Therefore, however many cylinders Steve can see will be the number of cylinders Steve thinks are on the table. However, sometimes there will be a cylinder that Steve cannot see. In such cases, he's wrong about the number of cylinders on the table. This means that Steve only actually knows the number of cylinders on the table when he can see the whole table.

After this background story, participants were shown a series of test stimuli to gauge whether they could accurately judge (i) the number of cylinders on the table, (ii) how many cylinders Steve thought were on the table, and (iii) whether Steve knew how many cylinders were on the table. This procedure was repeated, with feedback from the administrator, until participants could accurately attribute both belief and knowledge for cases in which Steve could and couldn't see the entire table. The aim of this pre-experimental procedure was to ensure participants shared a common sense in which they were to understand mental state terms, especially "knows," for purpose of the experiment. Additionally, great care was taken to provide all instructions in mental terms (e.g. "when Steve can see the whole table"), not non-mental terms (i.e. "when there is no partition"), in order to minimize the risk of participants applying a non-ToM heuristic to complete the ToM tasks (see §4.3 for evidence of this procedure's success).

The experiment utilized 25 different stimuli, each of which consisted of the same illustration of "Steve" sitting at a table. Stimuli differed in the number of cylinders on the table (0-3), the arrangement of those cylinders, and whether a partition blocked Steve's view of the back half of the table (see fig. 1). Each stimulus was presented for a maximum duration of 4 seconds, with a one second interstimulus interval consisting of a white fixation cross presented between every stimulus. Participant responses were recorded via the spacebar of a computer keyboard. Participants were instructed to use only one hand for responding, were permitted to select their dominate hand, and were instructed to otherwise keep still while completing the tasks. When a response was registered, the current stimulus presentation ended immediately, starting the routine for the next stimulus.

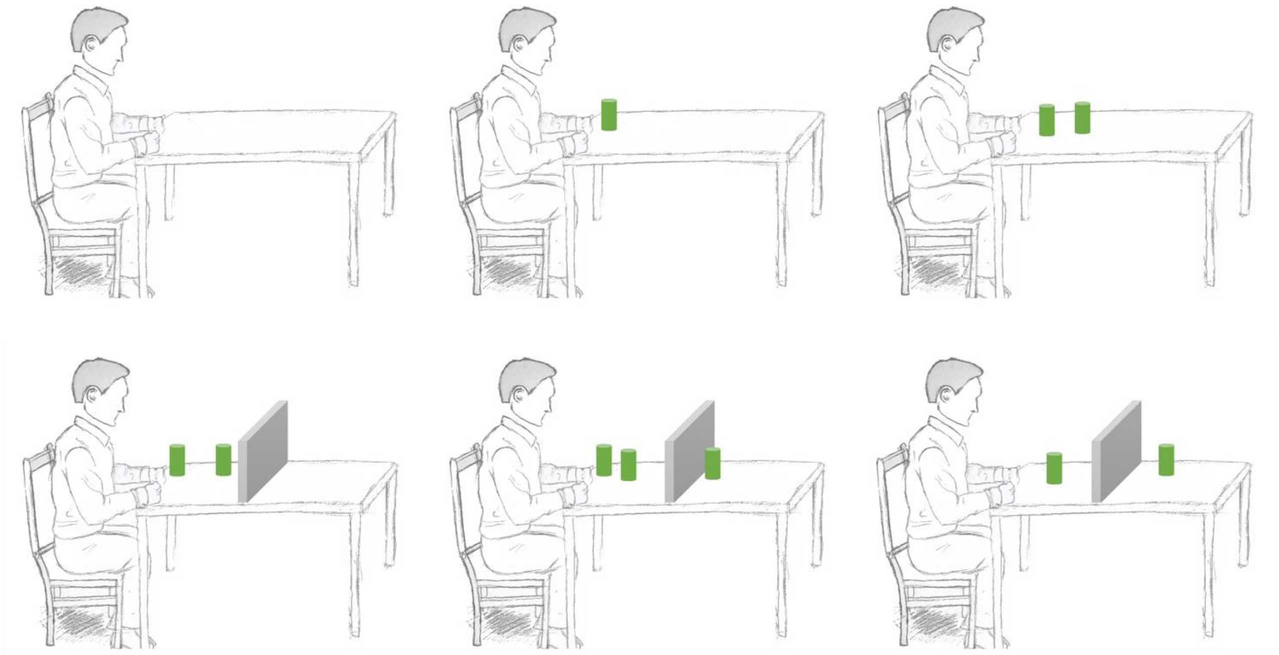

*Figure 1: Six representative stimuli, with (below) and without (above) partition*

The four experimental conditions of this study were defined by the task participants were instructed to complete. Each of these was a go/no go task that referred to two cylinders:

(T) The "truth" task elicited a judgement about the non-mental state of reality. Participants were instructed to press spacebar when there were exactly two cylinders on the table.

(B) The "belief" task elicited a belief attribution. Participants were instructed to press spacebar when Steve thought that there were exactly 2 cylinders on the table.

(TB) The "true belief" task elicited a composite judgement consisting of both a belief attribution and a judgement about the non-mental state of reality. Participants were instructed to press spacebar when Steve thought that there were two cylinders on the table and there were indeed two cylinders on the table.

(K) The "knowledge" task elicited a knowledge attribution. Participants were instructed to press spacebar when Steve actually knew that there were two cylinders on the table.

The key functional difference between the TB and K conditions is that Steve can never have knowledge of the number of cylinders on the table when there is a partition on the table. This is because the partition brings with it the easy possibility that Steve's representational states could fail to match reality, therefore precluding these states from displaying the stability in matching reality that characterizes knowledge states. Nonetheless, he might of course still have a true belief in the case that nothing is behind the partition.

The experiment was divided into 12 blocks of 50 trials each. Stimulus order was randomized within each block. Every block used the same set of stimuli, which entailed varying target stimulus probabilities between conditions (P(T) = P(B) = .44; P(TB) = .32; P(K) = .25). Each block consisted of only one task, and task instructions were confirmed verbally before the start of each block. 3 blocks were devoted to each task, so that participants completed a total of 150 trials for each task. Blocks for a given task were completed consecutively, so that that participants wouldn't move on to a new task until all three blocks for their present task were complete. Block order was partially counterbalanced between participants by alternating between four block sequences (K-TB-B-T; T-B-TB-K; B-T-K-TB; TB-K-T-B)

Stimuli were presented using a VPixx Technologies VIEWPixx /EEG display, and stimulus presentation was managed via PsychoPy3 (Peirce et al. 2019).

### *2.4 EEG Recording and Preprocessing*

EEG was recorded continuously using a Mega Electronics NeuroOne Tesla system (sampling rate: 500Hz). Participants were fitted with 30 passive electrodes arranged in the Easycap M3 (10/20) layout. Two EOG electrodes (1cm below/beside the left eye) were used to record eye movements. All electrodes were brought below 5kΩ and referenced to the nose during recording.

EEG data was preprocessed off-line using the MNE toolbox for Python (Gramfort et al. 2013). Each channel was re-referenced to the mean of all channels. Bad channels were identified manually and interpolated. Data was band-pass filtered from .1-40 Hz. Blink and saccade artifacts were identified via automated MNE functions and used in an independent component analysis (ICA) to help identify eye-movement components. The ICA decomposition was computed using an otherwise identical dataset filtered with a high-pass boundary of 1 Hz, and this decomposition was then applied to the primary (i.e. .1 Hz) dataset. The automated identification of eye-movement components in the ICA was verified manually to avoid over-removal of components, and the data was inspected visually before epoching.

ERP analysis was restricted to correct attributions/positive judgements (i.e. only the “go” trials): Stimulus-locked epochs were taken for every trial in which a participant registered a correct response, from 200 ms before stimulus onset to 800 ms after stimulus onset. Baseline correction to the 200 ms pre-stimulus interval was applied automatically during epoching. Epochs were visually inspected, and those displaying muscle artifacts were rejected manually (participant mean = 3, SD = 4.8).

### *2.5 Statistical Analysis*

Behavioral measures were analyzed using linear mixed-effects models. Analysis of reaction time data used task (T, B, TB, K) as the single fixed effect, with a random intercept for each subject. For error data, separate models were computed for miss rate (MR) and false-alarm rate (FAR). Both models used task as the single fixed effect with a random intercept for each subject.

Epoched EEG data was analyzed using one-factor repeated-measures ANOVAs of spatiotemporal data points (30 channels x 500 samples) of subject average ERPs, with the number of trials automatically equalized between conditions. The F-values for each data point were then transformed using threshold-free cluster enhancement (TFCE; E = .5, H = 2; Smith & Nichols 2009), and p-values for the TFCE-transformed values were obtained using a permutation test (N permutations = 10,000), taking the maximum statistic over all points to compensate for the multiple comparison problem. This analysis was conducted first over all four conditions (T, B, TB, and K; non-sphericity corrected using the Greenhouse-Geisser method) and then for each two-condition pair. Both the four-condition and pairwise analyses extended over the entire epoch (-200 to 800 ms). Effect sizes (Cohen’s d) for significant pairwise effects were calculated using differences between conditions at local maxima. All ERP analyses were performed using the MNE toolbox for Python (Gramfort et al. 2013). All statistical tests assume an alpha level of .05.

## 3. Results

### *3.1 Behavioral Results*

### *3.1.1 Reaction Time*

With K as the reference class, the model estimated an RT of 789 ms (t = 22.3; p < .001; 95% CI = [719, 860]). Significant differences were found between K and both T (beta = -86 ms; t = -12; p < .001; 95% CI = [-100, -72]) and TB (beta = +60 ms; t = 8; p < .001; 95% CI = [45, 75]). There was no significant difference between K and B (beta = +.1 ms; t = .02; p = .98; 95% CI = [-14, 14]). Observed effect sizes (Cohen’s d) were .42 for T-K and .27 for K-TB.

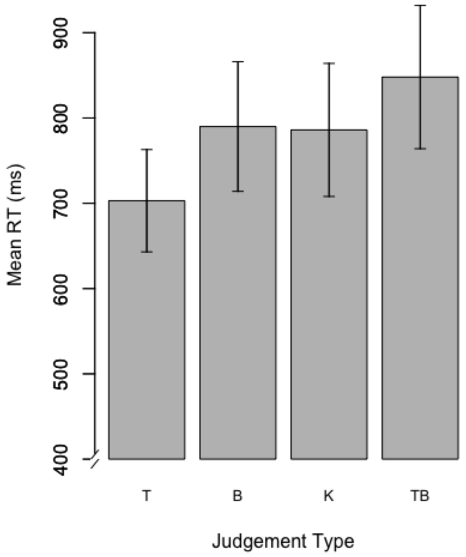

*Figure 2: Mean observed reaction times for each condition. Error bars are +/- 2 standard errors.*

### *3.1.2 Errors*

Error rates were very low (mean MR < .001; mean FAR = .001) No significant differences were found between conditions for either miss rates or false-alarm rates.

### *3.2 ERP Components*

### *3.2.1 All Conditions*

Observed differences between conditions (fig. 3) were indicative of four distinct clusters: (1) the P3b, (2) a posterior LSW, (3) a frontal LSW, and (4) a late, right-lateralized temporal effect.

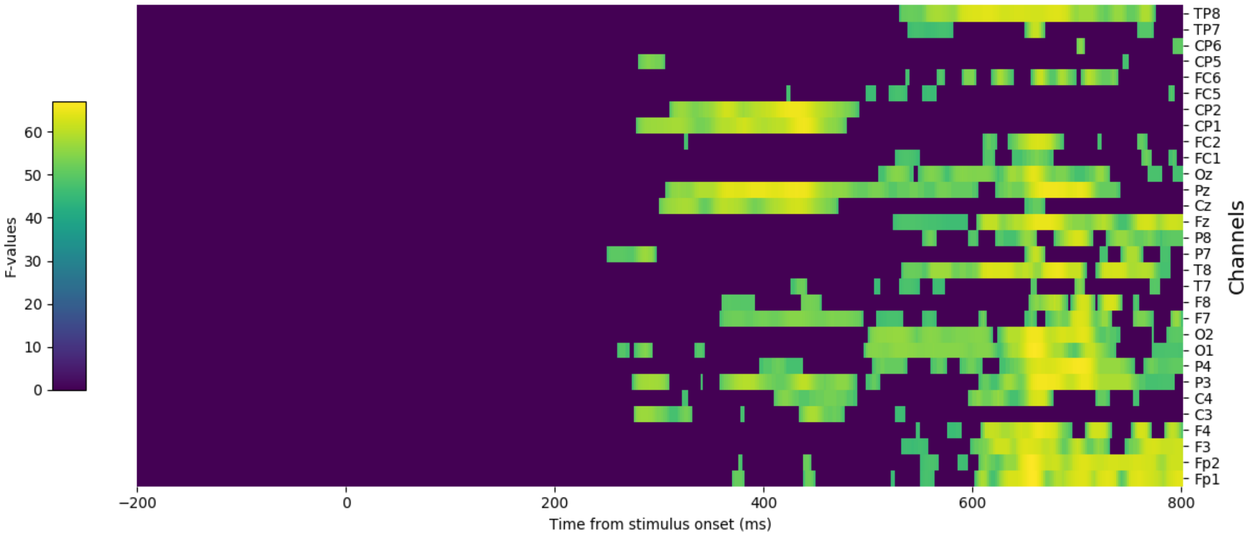

*Figure 3: Raster diagram depicting TFCE-transformed F-values observed over all conditions. Only the points for which the permutation test observed a significant difference (p < .05) are visible.*

Cluster #1: Channel Pz displayed a local maximum (p = .002) at 430 ms, which was most compatible with an effect at around 300-500 ms with a central centroparietal distribution. This effect was characterized by a greater amplitude for the T and K conditions than the B and TB conditions, and was maximal at the Pz, Cz, CP1, and CP2 electrodes (fig. 4). The latency, shape, and distribution of cluster #1 are all indicative of the parietal P3 component, P3b, (Polich 2011), and its onset is too early to be characteristic of the posterior LSW. On this basis, as well as the ubiquity of the P3 component for tasks requiring visual discrimination, we can reasonably understand cluster #1 to denote greater P3b amplitude for the T and K conditions than the B and TB conditions.

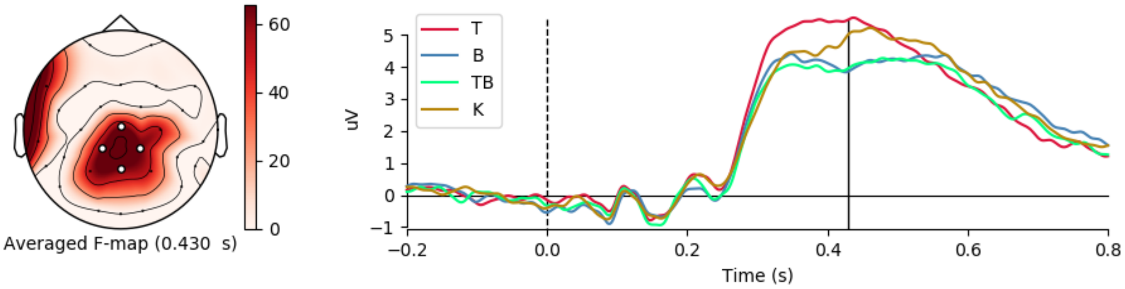

*Figure 4: Grand average ERPs for the four channels displaying the maximal P3b effect across all conditions (Pz, Cz, CP1, and CP2). The F-map depicts TFCE-transformed F-values, and the solid vertical line indicates the local maximum (channel Pz, p = .002).*

Cluster #2: Channel Pz displayed a local maximum (p = .002) at 674 ms, which was most compatible with an effect at around 600-800 ms with a central occipitoparietal distribution. This effect was characterized by a greater amplitude for the B, TB, and K conditions than the T condition, and was maximal at the Pz, P3, and P4 electrodes (fig. 5). The shape, latency, and distribution of cluster

#2 are indicative of the posterior LSW widely associated with mental state attribution. Accordingly, we might characterize cluster #2 as denoting greater posterior LSW amplitude for the conditions involving mental state attribution (B, TB, K) than the condition not requiring mental state attribution (T).

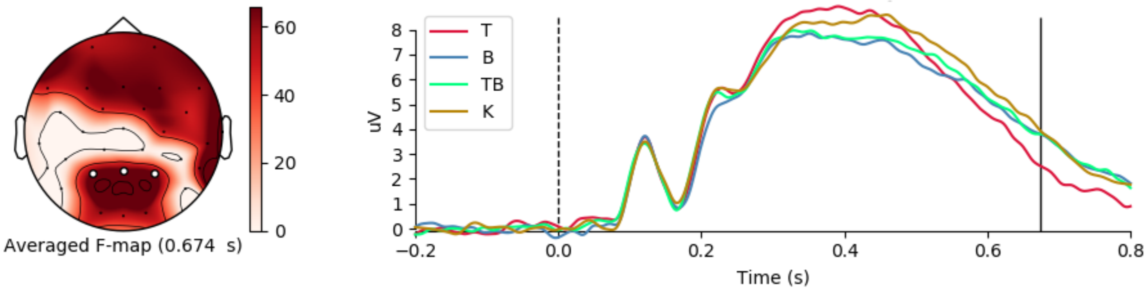

*Figure 5: Grand average ERPs for the four channels displaying the maximal posterior LSW effect across all conditions (Pz, P3, and P4). The F-map depicts TFCE-transformed F-values, and the solid vertical line indicates the local maximum (channel Pz, p = .002).*

Cluster #3: Channel Fz displayed a local maximum (p = .003) at 668 ms, which was most compatible with an effect at around 600-800 ms with a partially right-lateralized frontocentral distribution. This effect defies simple characterization, but displayed increasingly negative potentials for, in order, the T, B, TB, and K conditions. This effect was maximal at the Fz, Fp1, Fp2, F3, F4, and FC2 electrodes (fig. 6). The latency, shape, and distribution of cluster #3 are all indicative of the frontal LSW widely associated with mental state attribution.

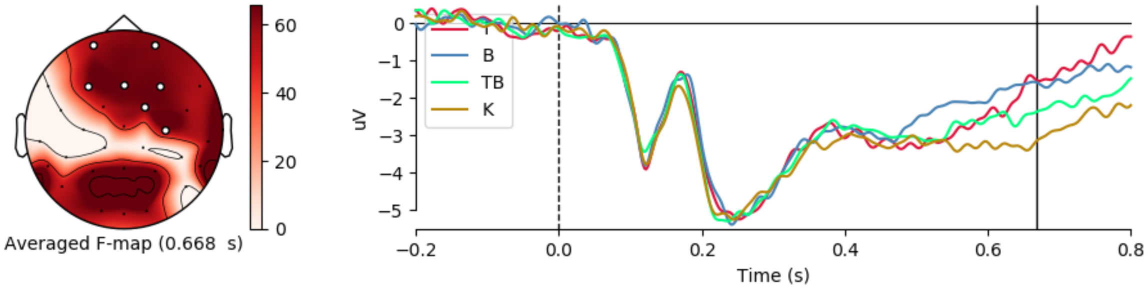

*Figure 6: Grand average ERPs for the seven channels displaying the maximal frontal LSW effect across all conditions (Fz, Fp1, Fp2, F3, F4, and FC2). The F-map depicts TFCE-transformed F-values, and the solid vertical line indicates the local maximum (channel Fz, p = .003).*

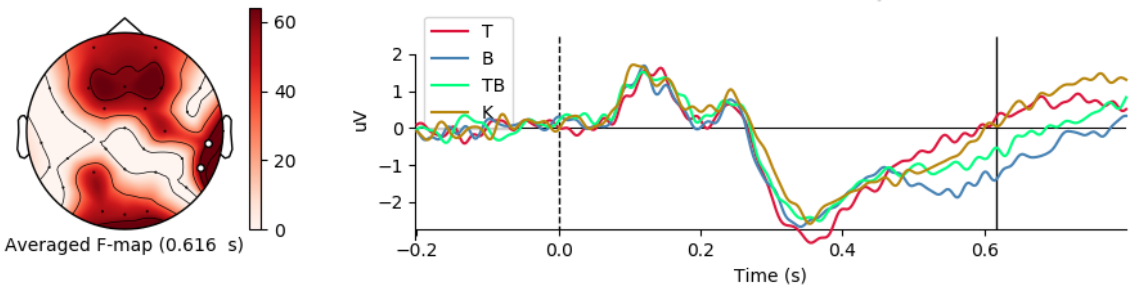

*Figure 7: Grand average ERPs for the two channels displaying the maximal right temporoparietal effect across all conditions (TP8 and T8) The F-map depicts TFCE-transformed F-values, and the solid vertical line indicates the local maximum (channel TP8, p = .003).*

Cluster #4: Channel TP8 displayed a local maximum (p = .003) at 616 ms, which was most compatible with an effect at around 500-800 ms with a right-lateralized temporoparietal distribution. This effect is roughly characterized by a more negative potential for the B and TB conditions vs. the T and K conditions, and was maximal at the TP8 and T8 electrodes (fig. 7).

### *3.2.2 Pairwise Comparisons*

Pairwise comparisons identified significant local maxima for every two-condition pair, with the exception of B-TB. The clusters most consistent with these local maxima are described below. It is especially important to observe that because K and B RTs don't significantly differ, these ERP differences are not confounded by RT differences.

Between the K and T conditions, two significant clusters were identified. Cluster #1: K amplitude > T amplitude at approximately 500-800 ms; central occipitoparietal distribution maximal at Pz, P3, P4, O1, and O2; local maximum at 662 ms (channel P3, p = .025); large effect size (d = .89). The latency, distribution, and shape of cluster #1 indicate a greater posterior LSW for K than T. Cluster #2: K amplitude > T amplitude at approximately 500-800 ms; central prefrontal distribution maximal at Fp1 and Fp2; local maximum at 656 ms (channel Fp1, p = .024); large effect size (d = .81). The latency, distribution, and shape of cluster #2 indicate a stronger (i.e. more negative) frontal LSW for K than T.

Between the K and B conditions, three significant clusters were identified. Cluster #1 (fig. 8): K amplitude > B amplitude at approximately 300-600 ms; central parietal distribution maximal at Pz; local maximum at 434 ms (channel Pz, p = .03); medium effect size (d = .66). The latency, distribution, and shape of cluster #1 indicate that it possibly reflects contributions from both the P3b and posterior LSW. As the latency of both the onset and local maximum are uncharacteristic of the posterior LSW, the best explanation for the earlier difference in cluster #1 is that it reflects a greater P3b amplitude for K than B. Cluster #2: K amplitude > B amplitude at approximately 500-800 ms; central frontal distribution maximal at Fz, F3, and F4; local maximum at 676 ms (channel F3; p = .02); large effect size (d = .80). The latency, distribution, and shape of cluster #2 indicate a stronger (more negative) frontal LSW for K than B. Cluster #3: B amplitude > K amplitude at approximately 500-800 ms; right temporoparietal distribution maximal at T8 and TP8; local maximum at 680 ms (channel T8, p = .016); large effect size (d = .88). The latency, distribution, and shape of this cluster indicate a stronger negative potential for B than K in the right temporoparietal effect identified in all-conditions cluster #4.

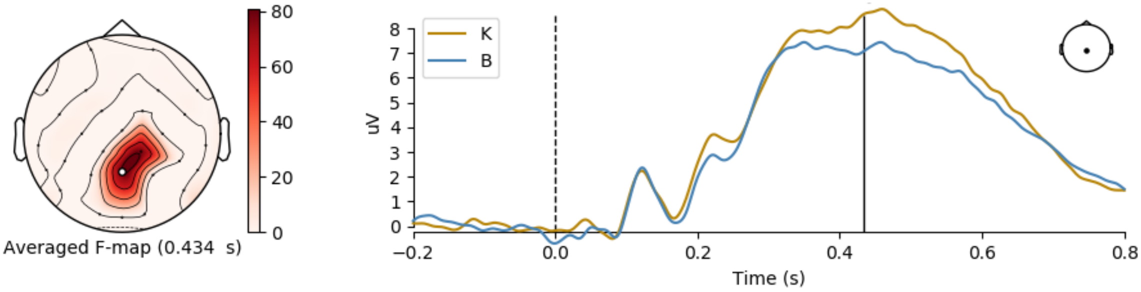

*Figure 8: Grand average ERPs for the channel displaying the maximal P3 effect for the K-B pairwise comparison (Pz) The F-map depicts TFCE-transformed F-values, and the solid vertical line indicates the local maximum (p = .03).*

Between the K and TB conditions, one significant cluster was identified: K amplitude > TB amplitude at approximately 300-600 ms; central parietal distribution maximal at Pz and CP1; local maximum at 444 ms (channel Pz, p = .04); large effect size (d = .81). The latency, distribution, and shape of this cluster indicate that it likely reflects contributions from both the P3b and LSW. As the latency of both the onset and local maximum are uncharacteristic of the LSW, the best explanation for the earlier difference is that it reflects a greater P3b amplitude for K than TB.

Between the T and B conditions, three significant clusters were identified: Cluster #1: T amplitude > B amplitude at approximately 300-500 ms; central centroparietal distribution maximal at Pz, Cz, CP1, CP2, and P4; local maximum at 378 ms (channel Pz, p = .004); large effect size (d = .84). The latency, distribution, and shape of this cluster all indicate a greater P3b amplitude for T than B. Cluster #2: B amplitude > T amplitude at approximately 500-800 ms; central parietal distribution maximal at Pz, P3, and P4; local maximum at 666 ms (channel Pz, p = .01); large effect size (d = .90). The latency, distribution, and shape of cluster #2 all indicate a greater posterior LSW for B than T. Cluster #3: B amplitude > T amplitude at approximately 500-800 ms; right temporal distribution maximal at TP8, T8, and F8; local maximum at 614 ms (channel TP8; p = .017); medium-to-large effect size (d = .77). The latency, distribution, and shape of this cluster indicate a stronger negative potential for B than T in the right temporoparietal effect identified in all-conditions cluster #4.

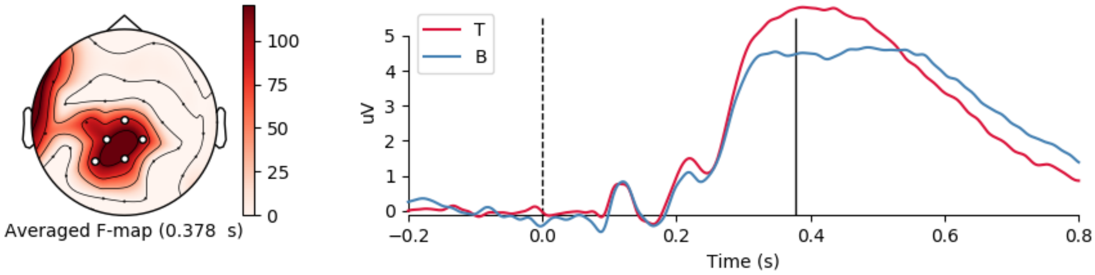

*Figure 9: Grand average ERPs for the channels displaying the maximal P3 effect for the T-B pairwise comparison (Pz, Cz, CP1, CP2, and P4) The F-map depicts TFCE-transformed F-values, and the solid vertical line indicates the local maximum (p = .004). Note also the transition to the posterior LSW effect around 500 ms.*

Between the T and TB conditions, two significant clusters we identified: Cluster #1: T amplitude > TB amplitude at approximately 300-500 ms; central centroparietal distribution maximal at Pz, Cz, CP1, and CP2; local maximum at 414 ms (channel Cz, p = .004); large effect size (d = .81). The latency, distribution, and shape of this cluster all indicate a greater P3b amplitude for T than TB. Cluster #2: TB amplitude > T amplitude at approximately 500-800 ms; central parietal distribution maximal at Pz, P3, and P4; local maximum at 680 ms (channel P4, p = .016); medium-to-large effect size (d = .75). The latency, distribution, and shape of cluster #2 all indicate a greater posterior LSW for TB than T.

No significant differences were found between the B and TB conditions.

These comparisons clarify the four significant differences between conditions identified in the all-condition ANOVA: (1) P3b amplitude for both T and K were greater than those for both B and TB, and no significant differences between T and K P3b amplitudes were identified. (2) The B, TB, and K conditions all demonstrated greater posterior LSW amplitude than the T condition. (3) The K condition displayed greater frontal LSW negativity than both the B and T conditions, which were the only significant differences observed for the effect. (4) The B condition displayed a more negative potential than both the K and T conditions in the right temporoparietal effect, which were the only significant differences observed for this effect.

| | *local max* | *p-vals* | *latency* | *channels* | *effect size* |
|---|---|---|---|---|---|
| ***P3b Cluster*** | **Pz @ 430 ms** | **.002** | **300-500 ms** | **Pz, Cz, CP1, CP2** | **-** |
| *K-B (+)* | Pz @ 434 ms | .03 | 300-600 ms | Pz | .66 |
| *K-TB (+)* | Pz @ 444 ms | .04 | 300-600 ms | Pz, CP1 | .81 |
| *T-B (+)* | Pz @ 378 ms | .004 | 300-500 ms | Pz, Cz, CP1, CP2, P4 | .84 |
| *T-TB (+)* | Cz @ 414 ms | .004 | 300-500 ms | Pz, Cz, CP1, CP2 | .81 |
| ***Posterior LSW*** | **Pz @ 674 ms** | **.002** | **600-800 ms** | **Pz, P3, P4** | **-** |
| *K-T (+)* | P3 @ 662 ms | .025 | 500-800 ms | Pz, P3, P4, O1, O2 | .89 |
| *B-T (+)* | Pz @ 666 ms | .01 | 500-800 ms | Pz, P3, P4 | .90 |
| *TB-T (+)* | P4 @ 680 ms | .016 | 500-800 ms | Pz, P3, P4 | .75 |
| ***Frontal LSW*** | **Fz @ 668 ms** | **.003** | **600-800 ms** | **Fz, Fp1, Fp2, F3, F4, FC2** | **-** |
| *K-T (-)* | Fp1 @ 656 ms | .024 | 500-800 ms | Fp1, Fp2 | .81 |
| *K-B (-)* | F3 @ 676 ms | .02 | 500-800 ms | Fz, F3, F4 | .80 |
| ***Right TP Effect*** | **TP8 @ 616 ms** | **.003** | **500-800 ms** | **TP8, T8** | **-** |
| *B-K (-)* | T8 @ 680 ms | .016 | 500-800 ms | TP8, T8 | .88 |
| *B-T (-)* | TP8 @ 614 ms | .017 | 500-800 ms | TP8, T8, F8 | .77 |

*Table 3: Summary of significant ERP findings. P-values and effect sizes (Cohen's d) are for local maxima. The latencies and channels provided are for clusters most consistent with observed local maxima. Order of pairwise comparisons indicates condition with the greater amplitude, and the following +/- indicates polarity (e.g. "K-T (-)" indicates a greater negative potential for K vs. T).*

## 4. Discussion

The primary aim of this study was to gain new insight into the neurocognitive mechanisms responsible for knowledge attribution, which had previously gone largely overlooked. We were especially interested in the question of whether knowledge is attributed (i) like a mental state, as is commonly assumed by psychologists, or (ii) a composite judgement partially constituted by belief attribution, as is commonly assumed by philosophers. The results of this study provide quite compelling behavioral and electrophysiological evidence in favor of the thesis that knowledge is attributed like a mental state. Moreover, these results further indicate that knowledge attribution, unlike belief attribution, does not recruit strong self-perspective inhibition during perspective taking. In addition to deepening our understanding of knowledge attribution, this study also sought to enhance our more general understanding of the electrophysiological correlates of Theory of Mind processes, especially the ubiquitous LSW components. As LSWs were observed during both knowledge and belief attribution, these results suggest that it is not correlated with self-perspective inhibition, likely reflecting some other ToM process, such as perspective taking.

### *4.1 Reaction Times: Contra the Composite Model*

The reaction time data collected in this study constitutes clear evidence against the composite model of knowledge attribution. Per the CM, belief attribution is a stage of knowledge attribution. This then predicts that, for otherwise comparable tasks, knowledge attribution should require more processing time than belief attribution. The failure to observe any significant difference in RTs between the K and B conditions is inconsistent with this prediction. Additionally, the data from the TB condition confirms the assumption that additional processing stages of a composite judgement results in longer RTs than simple mental state judgements. In sum, this behavioral data strongly suggests that our cognitive systems attribute knowledge like a mental state, not a composite of belief attribution with other judgements about non-mental reality. These findings are consistent with observations from experimental philosophy indicating that knowledge attributions can occur in the absence of belief attribution (Myers-Schulz & Schwitzgebel 2013; Murray et al. 2013).

A previous study from Phillips et al. (2017) also reported RTs inconsistent with the CM's prediction that RTs for judgements about knowledge should be greater for those about belief. However, unlike the present study, Phillips et al. reported lower RTs for knowledge than belief. This difference likely stems from multiple key differences between the two studies: (1) While the present study looked specifically at knowledge/belief attribution, Phillips et al. only reported data for knowledge/belief *evaluation*. That is, judgements both for and against knowledge/belief were analyzed together. (2) Perhaps more significantly, the Phillips et al. study used epistemic vignettes, as opposed to the simple ToM cartoons used in the present study, which undoubtedly entailed more complex judgements. We see evidence for more complex judgements in the much longer RTs observed by Phillips et al. (around 2.5 seconds), and this likely made the study better equipped to measure fine-grained differences between judgements about belief and knowledge. Nevertheless, all the empirical findings discussed here point in the same direction: Belief attribution is not a stage of knowledge attribution.

### *4.2 The P3b Component: Self-Perspective Inhibition*

This study observed statistically significant differences between tasks in the amplitudes of the P3b ERP component, with both the K and T conditions displaying greater P3b amplitudes than both the B and TB conditions. These results constitute evidence against the composite model of knowledge attribution, on which belief attribution is a stage of knowledge attribution. These results also constitute evidence against the belief-like model of knowledge attribution, on which self-perspective inhibition is strongly recruited during perspective taking. Instead, these results are consistent with the weakly inhibitory model of knowledge attribution, on which knowledge attribution is characterized by an absence of strong self-perspective inhibition during perspective taking. This result represents what is perhaps the most significant finding of this study. The following presents the rationale for this inhibitory interpretation in more detail.

The differences in P3b amplitude observed in this study are consistent with a straightforward explanation via resource allocation during P3 generation. As discussed above (§1.4), when the resource demands of additional, concurrent processes are sufficiently large during stimulus discrimination, this can limit the resources available for the inhibitory mechanism that produces the P3 component, thus resulting in a reduced P3 amplitude. We see this reflected clearly in the reduced P3 amplitudes for the belief attribution condition (B), confirming our assumption that (at least when engaged concurrently with perspective taking) the self-perspective inhibition recruited by belief attribution is strong enough to interfere with resource allocation to the P3 mechanism. This observation was facilitated in part by specific design features of this study, which increased the likelihood that we would observe this P3 effect for belief attribution. Most importantly, the block design meant that stimulus processing/recognition wasn't an antecedent condition for the recruitment of self-perspective inhibition. Participants knew to inhibit their own perspective even before stimuli were presented. Additionally, this study was careful to ensure that tasks would indeed lead participants to discriminate between stimuli on the basis of whether they depicted belief/true belief/knowledge states, maximizing the probability of producing this P3 effect.

Next, if we understand the composite judgements in the TB condition as computed serially (i.e., per the instructions, participants first attribute belief before assessing the truth value of that belief), we might also easily explain the reduced P3 amplitudes in the TB condition as reflecting the belief attribution stage of that task.

In contrast with the B and TB conditions, the truth condition (T) displayed a classically pronounced P3 component, with a significantly greater amplitude. This is easily explained by the fact that the task in the T condition is just a simple stimulus discrimination, without any ToM processing demands.

Finally, on the weakly inhibitory model of knowledge attribution, the resource allocation framework also provides a straightforward explanation for the observed greater P3 amplitude of the K condition when compared to both the B and TB conditions. If self-perspective inhibition is recruited at all, the process isn't demanding enough to significantly interfere with resource allocation during P3 generation. In contrast with belief, when we attribute knowledge, our cognitive systems aren't required to recruit an additional resource-intensive process. Accordingly, on the wIM, we shouldn't expect the same decreased P3 amplitudes associated with additional processing demands interfering with resource allocation to the P3 mechanism. This explanatory framework is not available on either the CM nor the BlM, both of which entail that knowledge attribution recruits the same strong self-perspective inhibition as belief attribution.

As a number of additional cognitive factors are known to modulate P3 amplitude, the possibility of an alternative explanation must be acknowledged. Beyond considerations of resource allocation, larger P3 amplitudes are also notably associated with (1) lower probability of target stimuli and (2) more difficult stimulus discrimination, and lower P3 amplitudes are often associated with (3) greater uncertainty of whether stimuli are targets (Polich 2011; Luck 2014). However, the results of this experiment are not consistent with a straightforward explanation via (1) or (2). Not only are the differences in target probability between conditions not especially large (.25-.44), but the T and B conditions, which displayed different P3b amplitudes, have equal target probabilities (.44). For this reason, the observed differences cannot be explained by a target probability effect alone. Similarly, the T condition cannot be reasonably thought to entail the most difficult discrimination, so the large P3b amplitude displayed is inconsistent with an explanation according to discrimination difficulty alone. Nevertheless, a more complex explanation involving both probability and difficultly cannot be ruled out. Additionally, while it cannot be ruled out that these findings are correlated with greater participant uncertainty during the B and TB conditions, there is nothing in the behavioral data that would indicate this.

### *4.3 The LSWs and Late Right Temporoparietal Effect: Theory of Mind Processing*

The later differences between conditions observed by this study provide a number of additional insights into how the brain processes mental state attributions.

The latency and functional characteristics of the frontal and posterior late effects observed in this study are indicative of the LSWs widely associated with ToM tasks (Sabbagh & Taylor 2001; Liu et al. 2004; Liu et al. 2009a/b; Zhang et al. 2009; McCleery et al. 2011; Meinhardt et al. 2011; Chen et al. 2012; Geangu et al. 2013). The B, TB, and K conditions all displayed significantly greater late posterior positivity than the T condition, and the K condition displayed significantly greater late frontal negativity than the T and B conditions.

McCleery et al. have suggested that differences in LSWs observed during false-belief tasks are reflections of differences in inhibitory control (2011, 12853). To the extent that this might be understood as inhibitory control of self-perspective information, the results of the present study indicate that such a hypothesis is unlikely. Per the P3b results, it appears that knowledge attribution recruits significantly less self-perspective inhibition than belief attribution. Were it the case that the posterior LSW corresponded with self-perspective inhibition, we would then expect greater posterior late positivity for B and TB than K. Similarly, were it the case that the frontal LSW corresponded with self-perspective inhibition, we would expect greater late negativity for B and TB than K. As we didn't observe either, it is likely that these LSWs are instead associated with some other ToM process or processes. Furthermore, because the same posterior late positivity was observed for both knowledge and belief attribution, it appears that the posterior LSW is a reflection of more general ToM processing, which isn't unique to belief attribution. However, as this increased late positivity has been previously observed to be absent for desire attribution (Liu et al. 2009a), we also have reason to think that this LSW doesn't reflect processing general to all mental state attribution. Conversely, the present findings indicate that the frontal LSW is a reflection of processing specific to knowledge attribution, at least to the extent that it is not shared by either belief attribution or judgements about non-mental reality.

Additionally, these LSW results indicate that it is unlikely that participants were employing a non-ToM heuristic strategy for the K task. Although participants were given instructions in strictly mental terms, it was possible that they might have employed a simple heuristic when completing the K condition, which avoided ToM processing entirely: press spacebar when there are two cylinders and no partition. The observation of a more positive posterior LSW for the K condition, matching the other conditions that involved judgements about mental states, indicates that this wasn't the case.

Nevertheless, there remains a good deal to be learned about the LSW components and their relation to Theory of Mind. The present study observed a more positive posterior slow wave in association with ToM tasks. However, while these results are similar to those presented by Liu et al. (2009a), a number of other studies have observed more negative (Sabbagh & Taylor 2000; Sabbagh et al. 2004; Liu et al. 2004) late posterior components associated with mental state attribution. As Meinhardt et al. have previously speculated, differences in LSW polarity likely derive from differences in tasks used to elicit them (2011, 74). Additionally, while several previous studies have reported differences in the frontal LSW between belief attributions and non-mental tasks (Sabbagh & Taylor 2000; Liu et al. 2004; Liu et al. 2009a/b), there were no significant differences observed between the present study's non-mental (T) and belief attribution (B) conditions. For these reasons, it is likely that the LSWs observed in the present study reflect neurocognitive mechanisms that are related but not identical to those of the LSWs observed by previous studies.

In addition to the frontal and posterior late slow waves, this study also observed a late effect displaying a right temporoparietal scalp distribution. This effect was characterized by a greater negative potential for the B condition vs. both the K and T conditions. These functional characteristics might be explained as differences in processing associated with mental content that doesn't match reality: As a number of fMRI studies have previously observed, the processing of mental states with false content is correlated with increased right temporoparietal activation vs. the processing of beliefs with true content (Sommer et al. 2007; Schuwerk et al. 2014; Bardi et al. 2017; Özdem et al. 2019). Because neither the T nor K conditions require the processing of mental content that doesn't match reality, but the B condition does, it is possible that the differences between these conditions and the B condition reflect this increased activation associated with processing false content. Given this, the findings of the present study then provide insight into the time course of the right temporoparietal processing associated specifically with false mental content. As this difference doesn't appear under after 500 ms, at least for the procedure of the present study, this indicates that it reflects a later process than the earlier self-perspective inhibition observed to interfere with the P3-generating mechanism.

Finally, it cannot be ruled out that the late differences observed here are at least partially attributable to differences in RT. However, as K and B display all these late ERP differences despite showing no differences in RT, and TB and B display no late ERP differences despite significant RT differences, these effects cannot be understood as solely a function of RT.

### *4.4 General Discussion*

Taken together, these results suggest the following characterization of the ToM processes that we utilize in attributing knowledge: At least in the case of simple perceptual knowledge, knowledge attribution is characterized by other perspective taking in the absence of strong self-perspective inhibition. The present study alone is insufficient to determine whether this means that knowledge attribution requires no self-perspective inhibition at all, or rather that it simply requires less inhibitory control over self-perspective information than belief attribution. Regardless of which of these ultimately proves correct, it is clear that self-perspective inhibition need not play the same strong role in knowledge attribution that it does in belief attribution. This reduced role for self-perspective inhibition reflects the relevance of self-perspective information during knowledge attribution. When determining whether some agent is in the state of knowing, we often appeal to self-perspective information about the state of reality, which frequently isn't available from the agent's perspective. Accordingly, inhibiting this information would be counterproductive to efficient knowledge attribution. This contrasts with belief attribution, on which such self-perspective information will never be relevant, and therefore might be categorically inhibited. In this way, we might understand that the weakly inhibitory model of knowledge attribution appears to track a key difference between the information required for successful knowledge vs. belief attribution.

In addition to the neurocognitive mechanisms of knowledge attribution, the findings of this study also contribute to a deeper understanding of the temporal dynamics of self-perspective inhibition itself. First, on the basis of LSW latency (600-800 ms), it has been previously suggested that inhibitory control is recruited later in belief state attribution (McCleery et al. 2011). However, the results of the present study indicate that self-perspective inhibition can engage early enough to contribute to differences in the P3b component (around 300ms). This early recruitment of self-perspective inhibition was aided by the design of this study, as participants knew whether they needed to inhibit their own perspective before stimuli were presented, and whether inhibition was required remained consistent for the duration of each 150-trial block sequence.

Finally, this study has a few key implications regarding the way philosophers think about knowledge and knowledge attribution. The most notable of these is that, at least in these cases of simple perceptual knowledge, knowledge is attributed like a mental state, not a composite. It is not trivial to assume that we can conclude on this basis that knowledge is indeed a mental state. However, such an approach has been defended by at least one prominent epistemologist (Nagel 2013, §4). Additionally, this appears to be the simplest available explanation for why treating knowledge like a mental state on the part of our ToM systems produces accurate judgements about knowledge: Knowledge is indeed a mental state. Given that this is the case, this would then mean that the bulk of epistemologists are mistaken in rejecting the idea that knowledge is a mental state. This would have significant ramifications for how epistemologists theorize about knowledge, as it is fundamentally at odds with the standard theoretical framework employed in the philosophy of knowledge (see Williamson 2000, ch. 1).

The cognitive differences between belief and knowledge attribution also call into question the philosophical practice of constructing theoretical arguments about knowledge attribution on the basis of the cognitive properties of belief attribution (see e.g. Nagel 2010; Gerken 2012). The danger of this practice is especially highlighted by Nagel's egocentric bias account of certain anomalous patterns of knowledge attribution (2010), which assumes analogous roles for self-perspective inhibition in belief and knowledge attribution. This study offers a clear demonstration that such an assumption, at least generally, is mistaken. While it could still be that there is some connection between belief and knowledge, on which certain cognitive properties of knowledge attribution can still be inferred from those of belief attribution, we can now understand that such a connection needs to be established before any such inference might safely occur. This underscores the importance of establishing the cognitive architecture of knowledge attribution empirically.

### *4.5 Limitations*

There are a couple limitations of this study that warrant further discussion, perhaps the most significant of which is the confounding of target probability with experimental condition. As previously mentioned (§4.2), P3 amplitude is known to be higher for stimuli with lower probabilities. We have a number of reasons to think that target probability was not the driving factor in the P3 differences observed in this study: (1) The B and TB conditions, which differed in target probability, displayed no significant P3 differences. (2) The T and B conditions, which did not differ in target probability, did display significant P3 differences. (3) The ordering of P3 amplitudes was generally not consistent with target stimulus probability. Nevertheless, keeping in mind that most of the observed P3 results cannot be explained as target probability effects, it cannot be completely ruled out that the moderate difference in target probability between the K and B conditions is responsible for at least some of the observed differences in P3 amplitude. While this is undoubtedly an unfortunate limitation of the present study, it should be noted that this limitation was not easily avoidable. First, we might note that in performing tasks like those used in this study, there is a non-trivial risk that participants might misunderstand mental state terms, especially "knowledge." It is not uncommon for the word "knowledge" to be used to non-literally to refer to states more closely resembling belief (e.g. "protagonist projection:" Holton 1997; Buckwalter 2014). Accordingly, it was decided that one crucial aspect of this study's design was that it should be determinable from behavioral data (e.g. false-alarm and miss rates) whether participants were actually understanding "knowledge" in the literal sense. Otherwise, it wouldn't be clear whether participants were even making knowledge attributions in the first place. Practically, this means there need to be stimuli that depict B and/or TB, which don't depict K. However, at least for simple scenarios that might be unambiguously depicted with ToM cartoons, {stimuli depicting K states} is a subset of {stimuli depicting TB states}, which is itself of course a subset of {stimuli depicting B states}. Indeed, most epistemologists working within the analytic project would even go so far as to say that there are no K states that are not also TB states. This nested structure of K, TB, and B states limited our options for presenting stimuli that depict B/TB without depicting K. While target probability could be equalized across conditions, this would require the use of different stimuli sets for different conditions. As this would then mean that any observed difference between conditions would be confounded with differences in stimuli, this was decided against. Instead, the design of this study reflected an effort to strike a balance between (i) the need for B- and TB-depicting stimuli that didn't depict K and (ii) the need to minimize target probability differences between conditions, all while using a single set of stimuli for all conditions.

A second limitation of this study is that it cannot definitively rule out alternative interpretations for the greater demand for neural resources during belief vs. knowledge attribution. For example, rather than a greater demand for inhibitory control, one might think that belief attribution entails the representation of non-obtaining states of reality to a greater extent than knowledge attribution. Nevertheless, while such an interpretation is well suited to the belief-like model of knowledge attribution, it is still inconsistent with the composite model: On the CM, a belief attribution is a stage of knowledge attribution, so there can be no processing inherent to belief attribution that doesn't then also occur for knowledge attribution. Thus, even if the differences in neural resource allocation indicated by this study do not derive from differences in self-perspective inhibition, they still support the conclusion that belief attribution is not a stage of knowledge attribution. However, evidence collected in this study casts doubt on such an alternative explanation for the P3 effect. The late time course of the right temporoparietal effect indicates that, at least on our specific experimental procedure, this representation of non-obtaining states of reality may not occur early enough to interfere with the P3-generating mechanism.

## 5. Conclusion

The findings of this study provide an important insight into the neurocognitive mechanisms that guide knowledge attribution. Most importantly, not only is belief attribution not a stage of knowledge attribution, but knowledge attribution doesn't require the strong self-perspective inhibition characteristic of belief attribution. These results allow us to better understand how our ToM systems generate attributions for different types of mental states, contributing to a fuller picture of how we think about the mental states of others. These results also raise a number of important questions in the philosophy of knowledge, especially regarding the reluctance of epistemologists to accept that knowledge is a mental state, as well as the philosophical practice of transposing the cognitive mechanisms of belief attribution directly onto questions about knowledge.

*Funding*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

*Acknowledgements*

I am singularly grateful to Henry Railo for his invaluable input over the course of this research. I am also thankful to a large number of other people who have facilitated its success, including: Teemu Rinne and everyone in human neuroscience at the University of Turku; the Faculty of Humanities at the University of Oulu; Laura Moreno Hernandez for assistance in participant recruitment; the attendees of the 2019 Swedish Congress of Philosophy in Umeå for their encouraging reception of the concept of this research; Mia Bricker for generously agreeing to draw the stimulus template; and Sanna Herrala for providing much-needed assistance proofreading. I also want to specifically thank the three anonymous reviewers from *Cognition* for their extremely valuable comments.

## References

Apperly, I. A., Samson, D., Chiavarino, C. & Humphreys, G. W. (2004). Frontal and Temporo-Parietal Lobe Contributions to Theory of Mind: Neuropsychological Evidence from a False-Belief Task with Reduced Language and Executive Demands. Journal of Cognitive Neuroscience, 16(10).

Bardi, L., Desmet, C., Nijhof, A., Wiersema, R., & Brass, M. (2017). Brain activation for spontaneous and explicit false belief tasks overlaps : new fMRI evidence on belief processing and violation of expectation. SOCIAL COGNITIVE AND AFFECTIVE NEUROSCIENCE, 12(3), 391–400.

Birch, S., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. Trends in Cognitive Sciences, 8(6), 255-260.

Bradford, E. E., Jentzsch, I. & Gomez, J. (2015). From self to social cognition: Theory of Mind mechanisms and their relation to Executive Functioning. Cognition, 138.

Buckwalter, W. (2014). Factive verbs and protagonist projection. Episteme, 11(4), 391–409.

Carrington, S., & Bailey, A. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. Human Brain Mapping, 30(8), 2313-2335.

Chen, L., Cheung, H., Szeto, C., Zhu, Z. & Wang, S. (2012). Do false belief and verb non-factivity share similar neural circuits? Neuroscience Letters, 510(1).

Covey, T., Shucard, J., & Shucard, D. (2017). Event-related brain potential indices of cognitive function and brain resource reallocation during working memory in patients with Multiple Sclerosis. Clinical Neurophysiology, 128(4), 604–621.

Duncan-Johnson, C., & Donchin, E. (1977). On Quantifying Surprise: The Variation of Event-Related Potentials With Subjective Probability. Psychophysiology, 14(5), 456–467.

Dumontheil, I., Apperly, I., & Blakemore, S. (2010). Online usage of theory of mind continues to develop in late adolescence. Developmental Science, 13(2), 331-338.

Ferguson, H. J., Apperly, I. & Cane, J. E. (2017). Eye tracking reveals the cost of switching between self and other perspectives in a visual perspective-taking task. The Quarterly Journal of Experimental Psychology, 70(8).

Fricker, E. (2009). Is Knowing a State of Mind? The Case Against. In Williamson on Knowledge. Oxford: Oxford University Press.

Geangu, E., Gibson, A., Kaduk, K. & Reid, V. M. (2013). The neural correlates of passively viewed sequences of true and false beliefs. Social Cognitive and Affective Neuroscience, 8(4).

Gerken, M. (2012). On the Cognitive Bases of Knowledge Ascriptions. In Knowledge Ascriptions. Oxford: Oxford University Press.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., . . . Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience, 7(7)

Hartwright, C. E., Apperly, I. A. & Hansen, P. C. (2012). Multiple roles for executive control in belief–desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. NeuroImage, 61(4).

Hartwright, C. E., Apperly, I. A. & Hansen, P. C. (2015). The special case of self-perspective inhibition in mental, but not non-mental, representation. Neuropsychologia, 67.

Heleven, E., & Van Overwalle, F. (2018). The neural basis of representing others' inner states. Current Opinion in Psychology, 23, 98-103.

Holton, R. (1997). Some telling examples: A reply to Tsohatzidis. Journal of Pragmatics, 28(5), 625–628.

Hyde, D. C., Simon, C. E., Ting, F. & Nikolaeva, J. I. (2018). Functional Organization of the Temporal-Parietal Junction for Theory of Mind in Preverbal Infants: A Near-Infrared Spectroscopy Study. The Journal of neuroscience: the official journal of the Society for Neuroscience, 38(18).

Ichikawa, J. and Steup, M., (2018) The Analysis of Knowledge. *The Stanford Encyclopedia of Philosophy (*Summer 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/>.

Isreal, J., Wickens, C., Chesney, G., & Donchin, E. (1980). The Event-Related Brain Potential as an Index of Display-Monitoring Workload. Human Factors: The Journal of Human Factors and Ergonomics Society, 22(2), 211–224.

Kramer, A., Wickens, C., & Donchin, E. (1983). An Analysis of the Processing Requirements of a Complex Perceptual-Motor Task. Human Factors: The Journal of Human Factors and Ergonomics Society, 25(6), 597–621.

Liu, A., Sabbagh, J., Gehring, M. & Wellman, M. (2004). Decoupling beliefs from reality in the brain: An ERP study of theory of mind. NeuroReport, 15(6).

Liu, D., Meltzoff, A., & Wellman, H. (2009a). Neural Correlates of Belief- and Desire-Reasoning. Child Development, 80(4), 1163–1171.

Liu, D., Sabbagh, M., Gehring, W., & Wellman, H. (2009b). Neural Correlates of Children's Theory of Mind Development. Child Development, 80(2), 318–326.

Luck, S. (2014). An introduction to the event-related potential technique (2nd ed.). Cambridge, US: Mit Press.

Mahy, C., Moses, L., & Pfeifer, J. (2014). How and where: Theory-of-mind in the brain. Developmental Cognitive Neuroscience, 9(C), 68-81.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods, 164(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

McCleery, J., Surtees, A., Graham, K., Richards, J., & Apperly, I. (2011). The neural and cognitive time course of theory of mind. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 31(36), 12849-12854.

Meinhardt, J., Sodian, B., Thoermer, C., Döhnel, K. & Sommer, M. (2011). True- and false-belief reasoning in children and adults: An event-related potential study of theory of mind. Developmental Cognitive Neuroscience, 1(1).

Murray, D., Sytsma, J., & Livengood, J. (2013). God knows (but does God believe?). Philosophical Studies, 166(1), 83–107.

Myers-Schulz, B., & Schwitzgebel, E. (2013). Knowing That P without Believing That P. Noûs, 47(2), 371–384.

Nagel, J. (2010). KNOWLEDGE ASCRIPTIONS AND THE PSYCHOLOGICAL CONSEQUENCES OF THINKING ABOUT ERROR. Philosophical Quarterly, 60(239), 286–306.

Nagel, J. (2013). Knowledge as a Mental State. In Oxford Studies in Epistemology Volume 4. Oxford: Oxford University Press.

Özdem, C., Brass, M., Schippers, A., Van Der Cruyssen, L. & Van Overwalle, F. (2019). The neural representation of mental beliefs held by two agents. Cognitive, affective & behavioral neuroscience, 19(6).

Peirce, J., Gray, J., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., . . . Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. Behavior Research Methods, 51(1).

Pergher, V., Wittevrongel, B., Tournoy, J., Schoenmakers, B., & Van Hulle, M. (2019). Mental workload of young and older adults gauged with ERPs and spectral power during N-Back task performance. Biological Psychology, 146, 107726.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology, 118(10), 2128–2148.

Polich, J. (2011). Neuropsychology of P300. In The Oxford Handbook of Event-Related Potential Components. Oxford: Oxford University Press.

Pritchard, D. (2005). Epistemic Luck. Oxford: Clarendon.

Royzman, E., Cassidy, K., & Baron, J. (2003). “I Know, You Know”: Epistemic Egocentrism in Children and Adults. Review of General Psychology, 7(1), 38-65.

Phillips, J., Knobe, J., Strickland, B., Armary, P., & Cushman, F. (2017). Knowledge before belief: Response-times indicate evaluations of knowledge prior to belief.

Sabbagh, M. A. & Taylor, M. (2000). Neural Correlates of Theory-of-Mind Reasoning: An Event-Related Potential Study. Psychological Science, 11(1).

Sabbagh, M., Moulson, M., & Harkness, K. (2004). Neural Correlates of Mental State Decoding in Human Adults: An Event-related Potential Study. Journal of Cognitive Neuroscience, 16(3), 415–426.

Samson, D., Apperly, I. A., Chiavarino, C. & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. Nature Neuroscience, 7(5).

Samson, D., Apperly, I. A., Kathirgamanathan, U. & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. Brain, 128(5)

Samson, D., Apperly, I. A. & Humphreys, G. W. (2007). Error analyses reveal contrasting deficits in “theory of mind”: Neuropsychological evidence from a 3-option false belief task. Neuropsychologia, 45(11)

Samson, D., Houthuys, S. & Humphreys, G. W. (2015). Self-perspective inhibition deficits cannot be explained by general executive control difficulties. Cortex, 70(C).

Shatz, M., Wellman, H., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. Cognition, 14(3), 301–321.

Schurz, M., Aichhorn, M., Martin, A., & Perner, J. (2013). Common brain areas engaged in false belief reasoning and visual perspective taking: A meta-analysis of functional brain imaging studies. Frontiers in Human Neuroscience, 7, 712.

Schurz, Matthias, Radua, Joaquim, Aichhorn, Markus, Richlan, Fabio, & Perner, Josef. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. Neuroscience and Biobehavioral Reviews, 42, 9.

Schurz, M., Kronbichler, M., Weissengruber, S., Surtees, A., Samson, D. & Perner, J. (2015). Clarifying the role of theory of mind areas during visual perspective taking: Issues of spontaneity and domain-specificity. NeuroImage, 117.

Schuwerk, T., Döhnel, K., Sodian, B., Keck, I., Rupprecht, R., & Sommer, M. (2014). Functional activity and effective connectivity of the posterior medial prefrontal cortex during processing of incongruent mental states. Human Brain Mapping, 35(7).

Smith, S., & Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage, 44(1), 83–98.

Smith, M. (2017). The cost of treating knowledge as a mental state. In Knowledge First: Approaches in Epistemology and Mind. Oxford: Oxford University Press.

Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. NeuroImage, 35(3), 1378–1384.

Turri, J. (2011). Mythology of the Factive. Logos & Episteme, 2(1), 141–150.

Turri, J. (2015). Evidence of factive norms of belief and decision. Synthese, 192(12), 4009–4030.

Turri, J. (2017). Knowledge Attributions and Behavioral Predictions. Cognitive Science, 41(8), 2253–2261.

Turri, J., Friedman, O., & Keefner, A. (2017). Knowledge central: A central role for knowledge attributions in social evaluations. The Quarterly Journal of Experimental Psychology, 70(3), 504–515.

Unger, P. (1968). An Analysis of Factual Knowledge. The Journal of Philosophy, 65(6), 157–170.

van Der Meer, L., Groenewold, N. A., Nolen, W. A., Pijnenborg, M. & Aleman, A. (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying Theory of Mind. NeuroImage, 56(4).

Watter, S., Geffen, G., & Geffen, L.. (2001). The n-back as a dual-task: P300 morphology under divided attention. Psychophysiology, 38(6), 998–1003.

Williamson, T. (2000). Knowledge and its limits. Oxford: Oxford University Press.

Zhang, T., Sha, W., Zheng, X., Ouyang, H. & Li, H. (2009). Inhibiting one's own knowledge in false belief reasoning: An ERP study. Neuroscience Letters, 467(3).