Knowledge and the Brain: Why the Knowledge-Centric Theory of Mind Program Needs Neuroscience

Adam Michael Bricker[a]

Forthcoming in Behavioral and Brain Science*s as a commentary on "Knowledge before belief" (Phillips et al. 2020).*

Abstract

The knowledge-centric Theory of Mind research program suggested by Phillips et al. stands to gain significant value by embracing a neurocognitive approach that takes full advantage of techniques like fMRI and EEG. This neurocognitive approach has already begun providing important insights into the mechanisms of knowledge attribution, insights which support the claim that it is more basic than belief attribution.

Main Text

The knowledge-centric approach advocated by Phillips et al. represents a welcome advancement in Theory of Mind research, and I am in complete agreement with this proposed shift in focus. My concern, however, is that Phillips et al. have overlooked an important source of evidence available to this emerging project—the *neuroscience* of knowledge attribution. Capable of providing insights even when undetectable in behavioral measures, as well as independent lines of converging evidence, hemodynamic (e.g. fMRI and fNIRS) and neurophysiological techniques (e.g. EEG and MEG) serve as powerful tools in Theory of Mind research. Crucially, neuroimaging has already begun to provide direct support for Phillips et al.'s central claim that knowledge attribution is more basic than belief attribution—belief attribution seems to demand neural resources that knowledge attribution does not (Bricker 2020). All this gives us compelling reason to think that the neuroscience of knowledge attribution has a vital role to play in the nascent knowledge-centric Theory of Mind research program.

It is not without good reason that neuroimaging techniques have been widely employed in the effort to understand our Theory of Mind systems (for overviews, see Carrington & Bailey 2009; Mahy et al. 2014; Schurz et al. 2014; Heleven & Van Overwalle 2018). A considerable amount of evidence indicates that the cognitive processes supporting human Theory of Mind capacities are both associated with identifiable neural correlates (see Heleven & Van Overwalle 2018) and distinct from more generalized executive function in the brain (see e.g. Hartwright et al. 2015; Samson et al. 2015; Bradford et al. 2020; Pacella et al. 2020). If the knowledge-centric Theory of Mind program is to achieve success comparable to that of its belief-centric counterpart, this observation is key. Mental state attributions are best understood not as cognitive, but rather *neuro*cognitive processes.

[a] Cologne Center for Contemporary Epistemology and the Kantian Tradition, University of Cologne. a.m.bricker@uni-koeln.de

The identifiable neural correlates of Theory of Mind processing enable neuroimaging techniques to provide additional lines of evidence that can converge with the findings of other methods. For example, fMRI studies indicating that the perspective taking and self-perspective inhibition components of Theory of Mind are largely supported by distinct regions in the brain (e.g. van Der Meer et al. 2011; Schuwerk et al. 2014; Hartwright et al. 2015; Özdem et al. 2019) have offered considerable support to the claim that these are indeed separate neurocognitive processes, which was initially suggested by Samson et al. primarily on the basis of lesion studies (2005).

Moreover, neuroimaging methods are especially valuable in their capacity to provide insights into Theory of Mind processing even when those insights aren't salient on behavioral measures. To take an example from fMRI, Hartwright et al. found differences in hemodynamic activity indicating that self-perspective inhibition during mental state attribution is distinct from inhibition during non-mental tasks, a finding that was not detectable in their behavioral data (2015). Taking a similar example from EEG, an N400 paradigm employed by Bradford et al. revealed initial egocentric processing during the attribution of false beliefs—even when those attributions were ultimately computed successfully (i.e. altercentrically)—providing key evidence that "egocentric processing is the default perspective for information integration" in such cases (2020, 276). Again, this evidence was not salient in their behavioral data

All this provides a general sense of the value of neuroimaging in Theory of Mind research. However, the neurocognitive findings most directly pertinent to the research program imagined by Phillips et al. come from my own EEG study (Bricker 2020). The first step in a broader research project dedicated to understanding the neurocognitive mechanisms of knowledge attribution, the design of this study was simple, with participants varyingly judging whether a cartoon character sitting at a table knew/believed that there were two cylinders on the table. This study provided two key results: (1) There were no significant difference in response time between belief attribution and knowledge attribution. (2) Differences in P3b amplitude indicated that the belief attribution tasks demanded a level of neural resources significantly greater than that of the knowledge attribution tasks, which is most likely explained by a greater demand for self-perspective inhibition during belief vs. knowledge attribution.

These findings are relevant to the account presented in the target article for at least two distinct reasons. First, these results provide additional evidence for the target article's central claim that knowledge attribution is at least as basic as belief attribution. As with the response time evidence discussed by Phillips et al. (§5.1), the idea that knowledge attribution relies on something like a belief attribution stage is inconsistent with the observation of comparable response times for belief and knowledge attribution tasks. We see something similar with the neurophysiological results, which indicate that belief attribution can entail processing demands that exceed those of knowledge attribution. This again suggests that knowledge attribution is the more basic of the two processes.

However, beyond simply providing further evidence that belief attribution does not come before knowledge attribution, the results of this study also illustrate why knowledge-centric Theory of Mind research works best when understood as a neurocognitive endeavor, highlighting both the advantages of neurocognitive techniques outlined above. Not only did the neurophysiological results of the study provide an additional line of evidence for the

conclusion suggested by behavioral measures, but these findings offered a further insight not salient on behavioral measures—Belief attribution appears to be a more resource-intensive process than knowledge attribution, likely due to differential demands for self-perspective inhibition.

While it is too early to speculate whether this knowledge-focused Theory of Mind research will ultimately attract the same amount of attention as its belief-centric counterpart, it is clear that neurocognitive techniques have a good deal to offer this emerging project. Through the integration of behavioral and neuroimaging methods with the characterization of knowledge states offered by epistemologists (see especially §2 of the target article; Bricker 2020, §1.1), we stand to make significant strides towards understanding the mechanisms underlying our judgements about knowledge, which are at present still largely unknown.

Funding Statement: This work was supported by Sven Bernecker's Alexander von Humboldt Professor grant.

Conflicts of Interest Statement: Conflicts of interest: none.

References:

Bradford, E., Brunsdon, V., Ferguson, H. (2020). The neural basis of belief-attribution across the lifespan: False-belief reasoning and the N400 effect. *Cortex*, 126, 265–280.

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

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.

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, 183–192.

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

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

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), 1433–1443.

Pacella, S. et al. (2020). Anosognosia for theory of mind deficits: A single case study and a review of the literature. *Neuropsychologia*, 148, 107641–107641.

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), 1102–1111.

Samson, D., Houthuys, S., & Humphreys, G. W. (2015). Self-perspective inhibition deficits cannot be explained by general executive control difficulties. *Cortex*, 70, 189–201.

Schurz, R., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34.

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), 2950–2965.

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).