Forthcoming in *Behavioral and Brain Sciences* as a comment on Phillips et al. ‘*Knowledge before belief*.’

Penultimate version. Please cite the published version.

Mikkel Gerken

***Representation and misrepresentation of knowledge***

**Abstract:** I argue for three points: First, evidence of the primacy of knowledge representation is not evidence of primacy of knowledge. Second, knowledge-oriented mindreading research should also focus on misrepresentations and biased representations of knowledge. Third, knowledge-oriented mindreading research must confront *the problem of the gold standard* that arises when disagreement about knowledge complicates the interpretation of empirical findings.

The target article (Phillips et al., this issue) provides converging evidence for assuming that representations of knowledge are more basic than representations of belief. While some findings that they take as evidence for representations of knowledge may perhaps be given deflationary interpretations, I am sympathetic to their broad *descriptive* conclusion about the primacy of knowledge representations. Likewise, I agree with their *methodological* conclusion that mindreading research should focus more on knowledge representation. Consequently, I will argue for three further points about knowledge-oriented mindreading research.

The first point is that it is fallacious to conclude that knowledge itself is primary from evidence that representations of knowledge are primary. Generally, it is fallacious to move from assumptions about the primacy of a *mental representation* to conclusions about the primacy of its *referent*. For example (from Gerken 2017b), it is a safe bet that representations of water are more basic than representations of hydrogen in terms of ontogenesis, phylogenesis, automatic processing etc. But this does not entail that the substance *water* is more basic or primary than *hydrogen*. Such a *representation-representandum* fallacy regarding representations of knowledge may occur in both epistemology and cognitive science. The fallacy is not committed in the target article, although its title – *Knowledge before belief* – might encourage it. So, to ensure that the surveyed evidence is used responsibly, it is important to warn against the *representation-representandum* fallacy. For example, it would be fallacious to take the surveyed evidence to motivate any knowledge-first program concerning knowledge rather than the concept of knowledge, the word ‘knowledge’ etc. (Williamson 2000).

The second point is that knowledge-oriented mindreading research should study misrepresentations and biased knowledge representations. The primacy of knowledge representations may be partly explained in terms of *bounded rationality*: Representations of knowledge are basic and prominent partly because they are cognitively “cheap” ways of representing complex epistemic matters in a manner that is accurate enough for many purposes. Insofar as representations of non-factive representations, which require one to keep track of both the mental representation and what it represents, are more cognitively taxing than representations of factive representations, cognitive bounds may partly explain the primacy of knowledge representations. However, bounded cognition involves biases.

The target article provides evidence that knowledge representations are automatically processed (§4.3). But it does not mention that automatic processing of primitive representations often exhibits signature biases (Saxe 2005; Apperly 2011). Likewise, misrepresentations of knowledge are not discussed. Given the aim of the target article, its focus on *successful* representations of knowledge is natural. However, this focus suggests a misleading picture of social cognition and overshadows the methodological upshot that knowledge-oriented mindreading research must also focus on misrepresentations and biases. Some research suggests that patterns of knowledge representations exhibit egocentric bias (Nagel 2008); focal bias (Gerken and Beebe 2015; Gerken et al. 2020); source-content bias (Turri 2015); subadditivity effects (Dinges 2018) etc. Signature biases of knowledge representations are important to study empirically because they are very consequential. For example, they may result in *discriminatory epistemic injustice* which occurs when someone is wronged specifically in her capacity as an epistemic subject (Fricker 2013: 1320; Gerken 2019). Since representations of knowledge play central roles in navigating social life, cases in which someone is wronged specifically in her capacity as a knower are especially harmful. Generally, given that knowledge representations are central to social cognition, it is important to empirically study their signature biases and the social ramifications thereof (Gerken 2017a; Spaulding 2018).

My third point is that knowledge-oriented mindreading research must confront a *problem of the gold standard*: Interpreting empirical data from tasks involving knowledge is often complicated because the gold standard response to the task is disputed. As noted, some researchers argue that particular patterns of knowledge ascriptions reveal a bias (Op. cit.). But others reject this and argue that these patterns reflect correct responses that illuminate the concept of knowledge, the word ‘knowledge’ or even knowledge itself (e.g., DeRose 2009; Knobe and Schaffer 2015; Stanley 2005). Some even suggest that to explain these patterns of knowledge ascriptions in terms of cognitive bias is to “explain away” the relevant evidence (Stanley 2005; DeRose 2009).

Presumably, the false-belief test is a widely employed experimental paradigm partly because of agreement about the gold standard response (e.g., saying that the agent will seek an object where she last saw it rather than at its new location in verbal false-belief tests (Wellman et al. 2001)). In contrast, the gold standard response is *part of what is investigated* in many mindreading tasks involving knowledge. This is simply a methodological challenge for research on knowledge representation and not a reason to stick with established experimental paradigms. Moreover, the problem of the gold standardis far from unique to research on knowledge representation, although it is pertinent to it because many aspects of knowledge are disputed. Minimal properties of knowledge, such as the four considered by Phillips et al. (§2), are good starting points. However, it is disputed how knowledge is related to luck, practical factors, actionability, competence etc. Knowledge-oriented mindreading research should study such relationships. But disputes over the gold standard response to tasks involving them constitute a methodological challenge in interpreting findings. Interestingly, the first two points mark specific methodological pitfalls. Given that some patterns of folk knowledge representation are biased (point two), moving too swiftly from findings about participants’ representations of knowledge to conclusions about whether they are correct would exemplify the *representation-representandum* fallacy (point one).

In sum, here are my three main points:

* Evidence of primacy of knowledge representation is not evidence of primacy of knowledge.
* Knowledge-oriented mindreading research should also focus on misrepresentations and biased representations of knowledge.
* Knowledge-oriented mindreading research must confront *the problem of the gold standard*.

These three points are compatible with the surveyed evidence and the main conclusions that Phillips et al. draw from it. But the points are not included in Phillips et al.’s conclusions. So, I wonder whether they agree with them or not.

**Acknowledgements:** Thanks to Kenneth Boyd, Uwe Peters and Shannon Spaulding. Funding: *Danmarks Frie Forskningsfund*, Grant Number: 8018-00053B.

***References***

Apperly, I. A. (2011). *Mindreaders. The Cognitive Basis of Theory of Mind*. Psychology Press.

DeRose, K. (2009). *The Case for Contextualism*. New York: Oxford University Press.

Dinges, A. (2018). Knowledge and availability. *Philosophical Psychology* 31 (4): 554-573.

Gerken, M. (2017a). *On Folk Epistemology. How we think and talk about knowledge*. Oxford: Oxford

 University Press.

Gerken, M. (2017b). Against Knowledge-First Epistemology. In Gordon and Jarvis Carter (ed.),

 *Knowledge-First Approaches in Epistemology and Mind*. Oxford University Press: 46-71.

Gerken, M., Alexander, J. Gonnerman, C. and Waterman, J. (2020). Salient Alternatives in

 Perspective. *Australasian Journal of Philosophy*.

Gerken, M., and Beebe, J. R. (2016). Knowledge in and out of contrast. *Noûs*, *50* (1): 133-164.

Nagel, J. (2008). Knowledge Ascriptions and the Psychological Consequences of Changing Stakes.

 *Australasian Journal of Philosophy*, 86: 279–94.

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences, 9* (4): 174-179.

Schaffer, J. and Knobe, J. (2012). Contrastive Knowledge Surveyed. *Noûs*, 46, (4): 675-708.

Spaulding, S. (2018). *How we understand others: Philosophy and social cognition*. Routledge.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford University Press.

Turri, J. (2015b). Skeptical Appeal: The Source-Content Bias. *Cognitive Science* 38 (5): 307-324.

Wellman, H., Cross, D. and Watson, J (2001). Meta-analysis of theory of mind development: The

 truth about false-belief. *Child Development*, 72 (3): 655-684.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.