# Theory of Mind in Context: Mental State Representations for Social Evaluation

Brandon M. Woo[1], Enda Tan[2], and J. Kiley Hamlin[2]

[1]Harvard University

[2]University of British Columbia

*This commentary, submitted to Behavioral and Brain Sciences, has not yet undergone editorial review.

**Target article:** Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., & Knobe, J. (2020). Knowledge before belief. *Behavioral and Brain Sciences*, 1-37.

Author Note

Correspondence concerning this commentary should be addressed to Brandon Woo.

Email: bmwoo@g.harvard.edu

**Abstract**

Whereas Phillips and colleagues argue that knowledge representations are more basic than belief representations, we argue that an accurate analysis of what is fundamental to theory of mind may depend crucially on the context in which mental state reasoning occurs. Specifically, we call for increased study of the developmental trajectory of mental state reasoning within socially evaluative contexts.

## Theory of Mind in Context: Mental State Representations for Social Evaluation

To support their argument that knowledge representations are more basic than belief representations, Phillips and colleagues (2020) draw on evidence suggestive that knowledge representations emerge earlier and more robustly in infancy than belief representations. They propose that knowledge representations, unlike belief representations, are developmentally privileged and "fundamental because they allow us to learn from others about [the true state of] the world." Here, we argue that learning the true state of the world is one, but not the only, function of theory of mind; therefore, it may be premature to conclude that knowledge is more fundamental than beliefs. Specifically, we argue for an increased focus on the role of mental state representations in contexts involving the evaluation of potential social partners (i.e., socially evaluative contexts).

Humans represent mental states not only to learn from others about the true state of the world, but also to learn about the character of potential social partners within it. Because humans must cooperate with each other to survive (Tomasello et al., 2012), we must be able to accurately assess potential social partners and determine whether they might cooperate with us in the future. Our theory of mind is crucial in this process, enabling us to distinguish, for instance, between an individual who intentionally poisoned someone's coffee and one who did so under the false belief that the poison was sugar (Young et al., 2007). Which individual would make a better social partner? Our ability to represent others' mental states, including not only what they know but also what they believe, is critical to evaluating others' actions and readily informs partner choice decisions (see Martin & Cushman, 2015).

Despite early developing motivations to form and maintain social relationships (see Raz & Saxe, 2020), the vast majority of studies on theory of mind development have not examined

mental state representations in socially evaluative contexts. Rather, infant theory of mind studies have almost exclusively assessed infants' expectations of a single, neutral agent who seeks to find an object. Phillips and colleagues include these studies as part of their evidence that knowledge representations are more basic than belief representations.

Without grounding studies of infants' mental state representations in contexts of social evaluation, however, it may be premature to form such conclusions. A large body of research has demonstrated that the context of a task may matter for false-belief reasoning as well as cognitive reasoning more broadly. For example, adults' cognitive reasoning is enhanced when tasks are framed as being about social contracts versus in non-social terms (Cosmides & Tooby, 1992). Similarly, a number of studies suggest that young children (who typically struggle in verbal tasks of false-belief understanding) may be better able to answer questions about false beliefs when agents act antisocially (Chandler et al., 1989; Tsoi et al., 2020; Wellman et al., 2001). Here, we explore the possibility that when infants engage in social evaluation, they show earlier capacities for mental state reasoning than previously assumed.

Take, for instance, studies assessing 3-month-old infants' understanding of others' goals. Past work has found that 3-month-olds do not readily represent the goals of agents' object-directed actions (Sommerville et al., 2005). And yet, 3-month-olds do appear to negatively evaluate agents who hinder others' goal pursuit: They selectively avoid looking at agents who steal a protagonist's ball (Hamlin & Wynn, 2011) and who prevent a protagonist's attempts to climb up a hill (Hamlin et al., 2010). These findings point to the possibility that 3-month-old infants may be more capable of representing others' (even unfulfilled) goals in socially evaluative versus non-evaluative contexts. Given these findings within the domain of goal

understanding, do socially evaluative contexts facilitate infants mental state representations more broadly?

Indeed, a growing number of studies have provided evidence that infants' social evaluations incorporate others' intentions and knowledge states by late in the first year (Hamlin, 2013; Hamlin et al., 2013; Woo et al., 2017), and recent work suggests that they may even incorporate others' false beliefs by 15 months. In Woo and Spelke (2020) 15-month-olds evaluated agents not on the basis of the objective consequences of their actions (whether they caused a protagonist to obtain a desired versus an undesired toy), but instead on the basis of whether or not the agents *believed* their actions would be helpful. That is, infants preferred an agent who directed a protagonist to a location where the agent had last seen a toy that they knew the protagonist desired (i.e., where the agent falsely believed the desired toy to be), over an agent who inadvertently directed the protagonist to the desired toy's actual location (i.e., where the agent falsely believed the desired toy was not). These findings replicated in two distinct testing contexts (in-person, online), and provide the first evidence that infants can reason about false beliefs in socially evaluative contexts.

Although these results clearly require replication by independent researchers, they stand in contrast with the mixed evidence that infants represent false beliefs in non-evaluative contexts (Poulin-Dubois et al., 2018). In light of research suggestive that the development of infants' goal understanding may also differ across socially evaluative and non-evaluative contexts, we call on future studies to systematically test whether the development of mental state representation differs across contexts. Given the importance of mental states, both factive and non-factive, for accurate social evaluation, infants may be more sensitive to what others believe earlier in

development in socially evaluative contexts than in the non-evaluative contexts of traditional false-belief tasks.

In sum, we argue that the study of theory of mind must consider the context in which mental state reasoning occurs. Although both knowledge and belief have major consequences for social evaluation, the vast majority of studies on infants' theory of mind to date have not examined infants' mental state representations in socially evaluative contexts, but instead in a comparatively inconsequential object search paradigm. By examining the development of belief representations in a wider range of contexts, we can better determine which mental states are fundamental to theory of mind.

**References**

Chandler, M., Fritz, A. S., & Hala, S. (1989). Small-scale deceit: Deception as a marker of two-, three-, and four-year-olds' early theories of mind. *Child Development*, *60*, 1263-1277.

Cosmides, L., Tooby, J. (1992). Cognitive adaptations for social exchange. In Barkow, J., Cosmides, L., Tooby, J. (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163-228). New York, NY: Oxford University Press.

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, *128*(3), 451-474.

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209-226.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30-39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, *13*(6), 923-929.

Martin, J. W., & Cushman, F. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors. *PLoS One*, *10*(4), e0125193.

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., ... & Knobe, J. (2020). Knowledge before belief. *Behavioral and Brain Sciences*, 1-37.

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., ... & Perner, J. (2018). Do infants understand false beliefs? We don't know yet–A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development*, *48*, 302-315.

Raz, G., & Saxe, R. (2020). Learning in infancy is active, endogenously motivated, and depends on the prefrontal cortices. *Annual Review of Developmental Psychology*, *2*.

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*(1), B1-B11.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology*, *53*, 673-692.

Tsoi, L., Hamlin, J. K., Waytz, A., Baron, A. S., & Young, L. *False belief understanding for negative versus positive interactions in children and adults*. https://moralitylab.bc.edu/wp-content/uploads/2020/09/Tsoi_MeanNiceAnne_children_adults.pdf

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

Woo, B. M., & Spelke, E. (2020). *Belief-based evaluations of helping in infancy*. PsyArXiv. https://https://doi.org/10.31234/osf.io/eczgp

Woo, B. M., Steckler, C. M., Le, D. T., & Hamlin, J. K. (2017). Social evaluation of intentional, truly accidental, and negligently accidental helpers and harmers by 10-month-old infants. *Cognition*, *168*, 154-163.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235-8240.