Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, 15, 626–636.

Spelke, E. S. (2022). *What babies know: Core knowledge and composition* (Vol. 1). Oxford University Press.

Tamis-LeMonda, C. S., Adolph, K. E., Lobo, S. A., Karasik, L. B., Dimitropoulou, K. A., & Ishak, S. (2008). When infants take mothers' advice: 18-month-olds integrate perceptual and social information to guide motor action. *Developmental Psychology*, 44, 734–746.

Teller, D. Y. (1979). The forced-choice preferential looking procedure: A psychophysical technique for use with human infants. *Infant Behavior and Development*, 2, 135–153.

Walk, R. D., & Gibson, E. J. (1961). A comparative and analytical study of visual depth perception. *Psychological Monographs*, 75(15, Whole No. 519), 1–44.

Core knowledge, visual illusions, and the discovery of the self

Marlene D. Berke* and Julian Jara-Ettinger

Department of Psychology, Yale University, New Haven, CT, USA
marlene.berke@yale.edu; julian.jara-ettinger@yale.edu
marleneberke.github.io; compdevlab.yale.edu

*Corresponding author.

Abstract

Why have core knowledge? Standard answers typically emphasize the difficulty of learning core knowledge from experience, or the benefits it confers for learning about the world. Here, we suggest a complementary reason: Core knowledge is critical for learning not just about the external world, but about the mind itself.

Spelke (2022) approaches core knowledge as an empirical question. And it is on these grounds that *What Babies Know* compellingly argues for innate representational systems of knowledge. But this approach leaves a central question unanswered: Why do we have core knowledge in the first place?

The answer is often taken as self-evident, and it is largely left implicit in Spelke's book. If humans evolved innate representational knowledge, it must be because these systems are (1) too difficult to learn from experience alone, and (2) critical for making the problem of learning about the world tractable. As Spelke puts it, humans develop a commonsense understanding of the world with much less data than machines require, and core knowledge (plus language) accounts for this difference.

Yet, these views center learning as a problem exclusively of building models of our environment. But core knowledge might be critical not only for learning about the world, but also for learning about our own minds. Metacognition – the ability to represent and build models of one's own mind – is widely agreed to be a foundational component of human intelligence, guiding how we learn (Flavell, 1979), shaping how we update our beliefs (Rollwage, Dolan, & Fleming, 2018), and possibly even forming the basis of self-awareness (Proust, 2013) and consciousness (Lau & Rosenthal, 2011; Peters, 2022). In some theories, it is uniquely human, and part of what separates humans from the rest of the animal kingdom (Carruthers, 2008). And yet, questions about how metacognitive representations might be learned or developed remain open. Core knowledge, we argue, might provide part of the answer.

Building a model of our own mind first requires that we distinguish mental representations that capture the external world from artifacts of how our mind works. This distinction can be far from clear-cut: Raw sensory data are processed by our perceptual systems with the goal of creating veridical representations (Berke, Walter-Terrill, Jara-Ettinger, & Scholl, 2022), but these computations can introduce (or fail to remove) distortions, sometimes leading to inaccurate representations of the outside world. In cases like those in Figure 1, our visual system's attempts to produce an accurate replica of the world end up, ironically, creating a compelling but incorrect representation. This is a deep challenge. Given the stream of mental representations built from sensory experience – and how phenomenologically compelling they all appear to us – how can we tell which parts reflect the external world and which parts ought to be mistrusted?

Consider how we might realize that the percepts in Figure 1 are illusions. As you approach the water on the highway ahead (Fig. 1A), it recedes and then vanishes as a function of your distance from it. If you see the artwork *Cercle et suite d'éclats* (by Felice Varini; Fig. 1B) in person, taking a few steps to the left or right would fragment and deform the floating circles, revealing that our viewpoint affects the objects' cohesion. And as we rove our eyes over the grid in Figure 1C, the circles at the intersections flicker between white and black, as though our eyes are somehow inducing action at a distance. In each of these cases, interpreting

Figure 1 (Berke and Jara-Ettinger). Examples of cases where perception provides compelling but incorrect representations of the world. (A) Mirage on the road. This is an atmospheric phenomenon where light bends upon encountering different densities of air. Perception fails to account for this distortion, leading to an illusory percept of water (in contrast to other light distortions that perception does account for and correct, such as color distortions because of shading, as in Adelson's checker shadow illusion; Adelson, 2001). (B) Felice Varini's (2009) *Cercle et suite d'éclats*, an art installation where curvatures were painted over a collection of houses, such that they appear as floating circles when seen from a particular viewpoint. (C) Scintillating grid illusion.

what we see against the backdrop of core knowledge enables us to realize that what we are seeing is not real.

When we identify that we are experiencing an illusion, we learn more than just that the world is not how it appears. We also learn about our own mind. Most directly, we learn to mistrust certain percepts, as when we realize that water on a highway can be safely ignored on a hot, sunny day. But the ability to identify illusions by comparing what we see against our core knowledge might also help us build more abstract models of our mind. Grounding experience on principles of reality helps us explicitly realize that mental representations are less trustworthy as a function of distance or lighting. Objects do not blur together when we take our glasses off; pencils don't break when dipped in water; and funhouse mirrors don't change our body.

These revelations are not limited to passive discoveries that happen when, by coincidence, we notice violations of core knowledge. As adults, we exploit core knowledge as a tool for reality testing. When we encounter something surprising, we carry out intuitive experiments over our mind: We are tempted to move our eyes and head to test how our visual experience changes. By having stable principles of how the world works, we can detect discrepancies between how the world seems (as determined by our perception) and how we know it ought to be (as determined by core knowledge), enabling us to build models of our own minds.

Our focus so far has been on object knowledge. Do other core knowledge systems also support learning about our minds? We believe this is the case. Consider how, when we lose our sense of direction, we do not wonder whether space has suddenly warped, but we instead attribute the lost sense of direction to a failure of our mind. Although such a realization might seem trivial, there are cases where core knowledge might help us make deeper discoveries about ourselves. Consider, for instance, the epistemic humility that comes knowing that we cannot always infer other people's mental states accurately. Such a realization might emerge from experiences where someone's behavior appears illogical, but we still hold on to the conviction that their actions must have resulted from a rational, goal-directed pursuit. These types of metacognitive representations not only help us understand ourselves, but they also make us better at navigating the world, at recognizing the limits of what we know, and at deciding when and how to explore so as to push those limits.

Does our proposal imply that all creatures that have core knowledge also have metacognition? This is unlikely: The process we propose requires core knowledge, but core knowledge alone is insufficient. At a minimum, an organism also needs (1) the capacity to instantiate representations over internal computations rather than over the external world – that is, metarepresentations – and (2) a learning algorithm that can build and refine metarepresentations by comparing experience against core knowledge (related work on artificial intelligence has shown proof-of-concept for such algorithms; Berke, Azerbayev, Belledonne, Tavares, & Jara-Ettinger, 2023). The prevalence of metacognition is therefore likely to be more restricted across species than core knowledge is.

Even if core knowledge did not evolve for the purpose of learning about our own minds, this does not make its ramifications for metacognition any less important. The use of core knowledge to learn metacognition may still be a major achievement in cognitive development. If we are correct, then core knowledge does not only play a pivotal role in learning about the external world, but also in learning about the internal world – how we think of ourselves, our mental lives, and who we are.

References

Adelson, E. H. (June 2001). On seeing stuff: The perception of materials by humans and machines. In B. E. Rogowitz, & T. N. Pappas (Eds.), *Human vision and electronic imaging VI* (Vol. 4299, pp. 1–12). SPIE.

Berke, M. D., Azerbayev, Z., Belledonne, M., Tavares, Z., & Jara-Ettinger, J. (2023). MetaCOG: Learning a metacognition to recover what objects are actually there. arXiv preprint arXiv:2110.03105.

Berke, M. D., Walter-Terrill, R., Jara-Ettinger, J., & Scholl, B. J. (2022). Flexible goals require that inflexible perceptual systems produce veridical representations: Implications for realism as revealed by evolutionary simulations. *Cognitive Science, 46*(10), e13195.

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language, 23*(1), 58–89.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373.

Peters, M. A. (2022). Towards characterizing the canonical computations generating phenomenal experience. *Neuroscience & Biobehavioral Reviews, 142*, 104903.

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness.* Oxford University Press.

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology, 28*(24), 4014–4021.

Spelke, E. S. (2022). *What babies know: Core knowledge and composition.* Oxford University Press.

Varini, F. (2009). Cercle et suite d'éclats. [Metallic paint on buildings]. Vercorin, Switzerland.

The role of language in transcending core knowledge

Susan Carey*

Psychology, Harvard University, MA, Cambridge, USA
scarey@wjh.harvard.edu

*Corresponding author.

Abstract

What Babies Know (*WBK*) argues that core knowledge has a unique place in cognitive architecture, between fully perceptual and fully conceptual systems of representation. Here I argue that *WBK*'s core knowledge is on the perception side of the perception/cognition divide. I discuss some implications of this conclusion for the roles language learning might play in transcending core knowledge.

Spelke's monumental *What Babies Know* (Spelke, 2022) provides evidence for six domains of core knowledge: Innate systems of abstract, structured, representations with long evolutionary histories. In *The Origin of Concepts* (*TOOC*: Carey, 2009), I drew on Spelke's work to provide evidence for core knowledge, and I welcome and endorse Spelke's extended and more nuanced characterization in *WBK*.