Studies in History and Philosophy of Science 92 (2022) 267–269

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Book Review

Representation in Cognitive Science by Nicholas Shea: But Is It Thinking? The Philosophy of Representation Meets Systems Neuroscience

There is a definition of cognition that can be succinctly expressed: It is computation over representations (Fodor, 1975), an idea that has since come under much attack and undergone several mutations but nevertheless has remained intact up to the present day (Milkowski, 2013; Ramsey, 2007). This formulation gives us what are often referred to as the computational and representational theories of mind (CTM and RTM). Most work on CTM and RTM, and how they relate to human thought, has been done by philosophers, psychologists and cognitive scientists. In contrast, neuroscience arguably began as the physiology of muscle and nerve, with a focus on sensorimotor behaviors in non-human model systems (e.g. the stretch reflex in the soleus muscle of the decerebrate cat). As of late, however, neuroscience has increasingly investigated processes that are characterized as cognitive – a rodent either planning a route in a maze (navigation) or making perceptual or value-based choices (decision-making). Notably, the word representation is used in neuroscience for all these cases: for sensorimotor behavior, for non-human animal cognition, and for full-blown human thinking. The question at hand is to ask is whether representation does useful conceptual work in all three cases and if it does, is it always referring to the same basic concept?

The idea of representation undergoes changes in vocabulary and meaning across philosophy, psychology, and neuroscience, morphing from intentional mental states to internal models to neural codes. In neuroscience, representation is used very loosely to refer to any mapping between either neural data or a brain region, and either the external world or behavior. For example, hand movement is represented in the hand knob region of primary motor cortex (Yousry et al., 1997), position is represented in the firing of muscle spindle afferents (Matthews, 1964), object features are represented in the ventral stream in the temporal cortex (DiCarlo et al., 2012), and body position in space is represented by place cells in the hippocampus (O'Keefe 1976). This use of the term representation is ubiquitous in the neuroscience literature and alludes merely to the fact it is possible to decode a stimulus or behavior-related variable from neural data. In parallel to the neuroscience literature, the philosophy of mind literature on representation is a veritable industry; some have even referred to it as the representation wars. For the most part, however, neuroscientists have gotten on with their experiments and theories without looking over at this mountain of contested philosophical material. This is regrettable because the thoughts philosophers have had about representation would almost certainly bring some much-needed nuance to the concept in neuroscience. This is especially apposite now given that, as already alluded to above, animal-model neuroscience has gotten into the cognition business. For example, the website at the Janelia Research Campus, under the section titled mechanistic cognitive neuroscience, states: "We aim to discover the circuit dynamics, network architectures, neuronal biophysics, synaptic rules, and molecular pathways that make cognition possible" (www.janelia.org). One of the model systems studied is the fruitfly *Drosophila melanogaster*. But is it really the case that we can learn about the neural basis of thinking by studying flies? It is precisely questions like this that benefit from philosophical consideration. Website entries like Janelia's tacitly imply that there are cognitive and *non*-cognitive behaviors, which means there must be a distinction between them; a distinction that can be better understood through an appreciation of philosophical considerations of representation. There are of course deflationary schools of thought that would rather either eschew the idea of representation altogether or at least diminish its sway. The dynamical systems approach (Van Gelder, 1995), embodiment (Fultot et al., 2019), and deep neural networks (Hasson et al., 2020) are examples of such positions. I will not discuss these here; one reason being that I consider, as does the author under review, such attempts to address cognition without the idea of representation a complete non-starter. That said, there *are* cases where the notion of representation is invoked when there is no need for it in neural explanations of behavior. That is to say, there *does* seem to be a boundary, albeit fuzzy in the hierarchy of the neuroaxis and/or behaviors, and it is only beyond this boundary that it becomes explanatorily useful to invoke the idea of representation for gaining insight into the link between brain and behavior. This brings us to Nicholas Shea's book *Representation in Cognitive Science* (RICS) (Shea, 2018).

RICS is exemplary in its rigorous and methodical arguments in support of a particular version of representational content, realized in physical vehicles, that is usefully posited as having a causal role in behavior. Such representations are to be contrasted with a 'factorized' account of behavior where a non-semantic causal chain will suffice. The example given of the latter is William Ramsey's description of the firing mechanism of a rifle from moving the finger on the trigger through the events that lead to the emergence of the bullet from the barrel (Ramsey, 2007). In this case, no useful explanatory work comes from a semantic relabeling of the process. The type of representation advocated for in the book refers to internal components of the nervous system that correlate or correspond structurally with distal features of the environment and are *flexibly* exploitable via algorithms in order to *robustly* complete a task function. The book is Shea meeting the *job description challenge* (Ramsey, 2007) for his notion of representation, naturalized with examples from neuroscience. The problem here is that the idea of representation he defends is too weak to be of use for a RTM. It appears to be an occupational hazard for many recent philosophers in this area to overly dilute their idea of representation in order to better engage with neuroscientific data. The assumption seems to be that this is the best bet for successful extrapolation at some point in the future to actual thinking. This book is an exemplar of ongoing projects to provide a naturalistic grounding for semantic content by aiming low to varying degrees. A hierarchy of choices offer themselves as entry points for this deflationary project: bacteria, sensorimotor systems, and intermediate cognition such as

https://doi.org/10.1016/j.shpsa.2021.05.014
Available online 5 July 2021

navigation. In a recent paper, Kolchinsky and Wolpert define semantic representation as "the information that a physical system has about its environment that is causally necessary for the system to maintain its own existence over time" (Kolchinsky & Wolpert, 2018). In this framework, a chemotactic bacterium swimming around in a nutrient solution has semantic information about its environment. In another attempt to naturalize the idea of representation as "stand-in" using experimental results from neuroscience, Piccinini (2020) has argued that sensory and motor areas contain structural representations that guide behavior. The argument seems to be that they are representations because they are a processed form of the raw sensory data; an abstraction that is needed for the organism to behave. Note that there is no qualitative difference here from the previous claim for what a bacterium is doing. In these two cases the posited representation is occurring simultaneously with the presence of the stimulus or occurrence of an action. The idea, central to cognitive representations, that they can operate uncoupled from the external world is not invoked. Indeed, such detachment, whereby a representation can guide stimulus-free performance, what Orlandi calls "coordinating with absence" (Orlandi, 2021), gets deemphasized in these examples and in Shea's book. To be fair, Piccinini does address uncoupling but relies on the examples of working memory and procedural memory (Piccinini, 2020). These transient stand-ins for a stimulus or a look-up table hardly meet the requirements for the rich intentional representations that would be required for cognition – they are not models of the world that can be flexibly used for simulation. Shea does allude to detachment in the context of place cells for navigation in rodents, when he writes that the "firing of the place cells is taken offline, that is, it is no longer directly driven by input about the animal's current location." This is very important because sequences of firing of place cells are posited to be used for selecting between possible routes in a maze, referred to by Shea as "preplay". This would be closer to a real representation, a cognitive map, that stands in for the real maze. To date, however, there is no compelling causal evidence that such replay (a better term to use for the host of such hippocampal phenomena) is *used* by the rodent for planning; the causal claim for vicarious trial and error is premature. Nevertheless, this offline, detachable representation *would* seem to be more relevant for extrapolation to mental representation but it is not the kind pushed for in Shea's book.

Shea, like Piccinini (2020) more recently, in his treatment of representation eschews discussion of *mental* representations. He opens the book referring to the mystery of thinking but then quickly sidesteps doxastic states and focuses instead on what he calls subpersonal representations. The justification provided for looking at simpler neural representations is predictable. From Piccinini we are told that they serve as "building blocks" for intentional mental states (Piccinini, 2020), and Shea tells us they will "prove a useful staging post on the way to tackling the more complex cases." I for one am unconvinced about this hopeful extrapolation. Shea's core idea can arguably be found in the form of a diagram on page 202, showing that there are internal representations that allow for invariance of behavioral goal or outcome despite different proximal inputs and movement outputs. Shea refers to these as vehicles of content that 'bridge" between inputs and outputs in a more general way that capture "real patterns in organism-world relations". The idea seems to be that having internal states interposed between inputs and outputs leads to a more abstract representation at the goal level, which is more parsimonious than positing many individual causal chains for each input-output pair. Shea argues that such bridging allows for semantic over non-semantic explanations of behavior and thus paves the way for full-blown psychological explanations. The deep problem here is that Shea either does not consider or perhaps is not familiar with the notion of flexible control policies, which can operate without any need to invoke internal models or simulation. The temptation is to invoke representations whenever an intelligent flexible behavior is observed but it is not necessary. To illustrate the point, I will give the example of the pithed frog. In the mid-nineteenth century, the German physiologist Eduard Pflüger conducted experiments in which frogs had their spinal cord severed and brain removed (pithing) (brewminate.com). They were then suspended from a hook and a piece of paper dipped in acetic acid was applied to their torso. The observation was that the frog was able to accurately wipe away the acid with its back foot. If that foot was amputated it wiped the acid away with its other foot, without reflexively activating its useless stump. These experiments were a shock to many as they seemed to show purposive reflexive behavior. The thing to notice is that the frog's behavior fits exactly into Shea's bridging scheme. The inputs can vary – the acetic acid can be placed in various locations and the outputs can vary – even different legs can be used. The goal, however, to wipe away the irritant remains invariant. Such intelligent flexible reflexes have since been described innumerable times, including in humans (Krakauer, 2019). The critical point is that there is no need to invoke a representation, just causally relevant external and internal states that can be input into a flexible control policy (Haith & Krakauer, 2013). Representation for cognitive purposes will need to extend beyond this impoverished notion – thinking is going to require a type of representation qualitatively different from the kinds being used by the pithed frog; the kinds posited by Shea. Instead one needs to appeal at the very least to the idea of internal models – an overt simulation of the external world that is actually used. This argument has played out in the motor control literature, some of which Shea cites, but it has become apparent, for example, that there is no need to invoke forward models for motor planning or adaptation (Hadjiosif et al., 2021). Similarly, a non-model-based formulation, the successor representation, can explain hippocampal-dependent navigation in rodents (Momennejad et al., 2017). Indeed, it is surprisingly difficult to definitively demonstrate model-based behavior in non-human animals and in humans it appears associated with explicit knowledge (Castro-Rodrigues et al., 2020). A deep neural network case for a non-representational account of most behavior has recently been made (Hasson et al., 2020). Notably, however, in this paper it is conceded that the nonrepresentational account will not work for cognition. Thinking, broadly captured by Kahneman's (2013) system 2, is not going to be explained by intelligent reflexes, i.e., ever more elaborate system 1 "representations", which are more accurately referred to as control policies. Ramsey intuited this distinction back in his 2007 book (Ramsey, 2007). Yoshua Bengio (2019), a pioneer of deep learning, has been very vocal of late that current AI lacks system 2 and needs to try and replicate it. There is an important lesson in the fact that current AI and Shea's formulation from within neuroscience both fall short of cognition in a similar way. To conclude, Shea correctly invokes the need for an idea of representation for thinking, but then zeroes in on a form of it that does not have the requisite world-model characteristics. Shea, almost inadvertently, has instead discovered flexible control policies but these by definition do not need a rich notion of representation. Alas, we are no closer to naturalizing thinking.

References

Bengio, Y. (2019). *From System 1 Deep Learning to System 2 Deep Learning*. NeurIPS. https://www.youtube.com/watch?v=FtUbMG3rlFs.

Castro-Rodrigues, P., Akam, T., Snorasson, I., Camacho, M., Paixão, V., Barahona-Corrêa, J. B., … Oliveira-Maia, A. (2020). Explicit knowledge of task structure is the primary determinant of human model-based action. *MedRxiv*. https://www.medrxiv.org/content/10.1101/2020.09.06.20189241v1.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434.

Fodor, J. (1975). *The language of thought*. New York: Thomas Y Crowell.

Fultot, M., Adrian Frazier, P., Turvey, M. T., & Carello, C. (2019). What are nervous systems for? *Ecological Psychology, 31*(3), 218–234.

Hadjiosif, A. M., Krakauer, J. W., & Haith, A. M. (2021). Did we get sensorimotor adaptation wrong? Implicit adaptation as direct policy updating rather than forward-model-based learning. *Journal of Neuroscience, 41*(12), 2747–2761.

Haith, A. M., & Krakauer, J. W. (2013). Model-based and model-free mechanisms of human motor learning. In *Progress in motor control* (pp. 1–21). New York: Springer.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron, 105*(3), 416–434.

https://brewminate.com/the-curious-case-of-the-decapitated-frog/

Kahneman, D. (2013). *Thinking, fast and slow*. New York: Farrar, Strauss and Giroux.

Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus, 8*(6), 20180041.

Krakauer, J. W. (2019). The intelligent reflex. *Philosophical Psychology, 32*(5), 823–831.
Matthews, P. B. (1964). Muscle Spindles and their motor control. *Physiological Reviews, 44*, 219–288.
Milkowski, M. (2013). *Explaining the computational mind*. Cambridge: MIT Press.
Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behavior*, (9), 680–692.
O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental Neurology, 78*, 109.
Orlandi, N. (2021). Representing as coordinating with absence. In *What are mental representations?* (pp. 101–134). New York: Oxford University Press.
Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford: Oxford University Press.
Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.
Shea, N. (2018). *Representation in cognitive neuroscience*. Oxford: Oxford University Press.
Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy, 92*, 345–381.
Yousry, T. A., Schmid, U. D., Alkadhi, H., Schmidt, D., Peraud, A., Buettner, A., & Winkler, P. (1997). Localization of the motor hand area to a knob on the precentral gyrus. A new landmark. *Brain, 120*, 141–157.

John W. Krakauer
The Johns Hopkins University School of Medicine, United States
E-mail address: jkrakau1@jhmi.edu.