4-1-2022

Some dilemmas for an account of neural representation: A reply to Poldrack

Michael L. Anderson
Western University, mljanderson@gmail.com

Heather Champion
Western University

Some dilemmas for an account of neural representation: A reply to Poldrack

Michael L. Anderson[1,2,3] Heather Champion[1,2]

1. Department of Philosophy
2. Rotman Institute of Philosophy
3. Brain and Mind Institute

University of Western Ontario, London, ON CANADA N6A 3K7

mande54@uwo.ca, hchampi2@uwo.ca

Abstract

"The physics of representation" (Poldrack 2020) aims to (1) define the word "representation" as used in the neurosciences, (2) argue that such representations as described in neuroscience are related to and usefully illuminated by the representations generated by modern neural networks, and (3) establish that these entities are "representations in good standing". We suggest that Poldrack succeeds in (1), exposes some tensions between the broad use of the term in neuroscience and the narrower class of entities that he identifies in the end, and between the meaning of "representation" in neuroscience and in psychology in (2), and fails in (3). This results in some hard choices: give up on the broad scope of the term in neuroscience (and thereby potentially opening a gap between psychology and neuroscience) or continue to embrace the broad, psychologically inflected sense of the term, and deny the entities generated by neural nets (and the brain) are representations in the relevant sense.

1. Introduction

As the title, "The Physics of Representation" suggests (Poldrack, 2020), this is a very ambitious paper that, ultimately, seeks to establish the claim that representations are *necessary* for intelligent behavior—a universal feature of behaving organisms. Although we won't discuss this bold claim in much detail, we highlight it here because we think that, while it may well be true in some form—for instance, dimensional reduction, finding structure in signals, is at least a plausible candidate for a universal feature of perception—Poldrack's overall analysis does not support the robust gloss he gives the claim.

To show why, we focus instead on three more modest goals of the paper: (1) to describe how the term "representation" is used in neuroscience and characterize its referent; (2) to offer an

existence proof of such representations in an Artificial Neural Network (ANN); and (3) to argue that such representations are "legitimate representations" according to a widely endorsed philosophical account. We argue that he succeeds in (1); exposes some unwanted and unremarked dilemmas for himself in (2); and fails in (3). The failure of claim (3) would appear to require embracing one horn of the dilemmas exposed in (2), with the interesting consequence that the existence of the representations described in (1) have yet to be definitively established, and the representations that *have* been established are of little use to psychology.

A note on vocabulary: since "representation" is the central contested term at play here, we will adopt the following conventions: "representation" will remain the generic term, its meaning to be determined, although not stipulated in this response. "Neural representations" (NRs) will be used to designate the entities *actually* identified in the neurosciences as carrying information for cognitive/behavioral purposes. "Mental representations" (MRs) will refer to the entities posited in the representational/computational theory of mind and broadly referenced in psychology. And "Artificial representations" (ARs) will refer to the information carrying entities identified in artificial neural networks. A central question for this paper is the relation between "representations", ARs, NRs, and MRs. There is a tension, as will become apparent, between the broad deployment of the term "representation" in neuroscience, and the entities designated as NRs. We will use the generic term "representation" when we are seeking to emulate this broad use. (Note these conventions are idiosyncratic to this current paper; we adopt them only to help maintain clarity.)

2. "Representations" in neuroscience

Poldrack's account of representations centers on the undeniable fact that the term is in common use in the neurosciences, used "to describe the systematic empirical relationships that are often found to exist between neural activity and features of the external world" (Poldrack, 2020: 1-2). An important question raised by this usage is what the relationship might be between the entities identified in the neurosciences and the *mental representations* posited by the Representational Theory of Mind (RTM), a position these days largely co-extensive with the Computational Theory of Mind (CTM) widely assumed (and endorsed) by psychologists and neuroscientists (Putnam 1967; Fodor 1975; 1981; Gallistel & King 2009). Indeed, the widespread assumption that the brain is a certain sort of computational system is what licenses the analogy with Artificial Neural Networks (ANNs) that drives an important aspect of Poldrack's treatment.

To explore this question, Poldrack attempts to flesh out the nature of neuroscience representations (NRs) in the following way:

> The presence of neural responses that are organized in a way that is structurally isomorphic with the external world has been known for more than a century… By "structurally isomorphic" I mean here that there is a systematic relationship between the activity of neurons and the structural (usually spatiotemporal) features of the world, such that the larger-scale organization of neural activity maps onto the structure of the world at the relevant scale. (Poldrack 2020: 2).

Poldrack is of course well aware that a *mere* "systematic empirical relationship" between neural activity and features of the world isn't going to support the case that NRs are "legitimate" representations, meeting Ramsey's (2007) "job description challenge", i.e., possessing the distinctive characteristics that qualify them for their special explanatory role in cognitive systems. After all, natural representations and indicators, which are ubiquitous in nature (tree rings, smoke), bear systematic empirical relationships to other things (tree age, fire), but are not generally taken to meet the challenge (although see Rupert 2018). In contrast (if there is a contrast, see Nirshberg & Shapiro 2021) structural representations generally *are* taken to be good candidates for legitimate representations because the preserved resemblance—the isomorphism—between the representation and its referent supports things like reasoning about properties of the target (Gladziejewski & Milkowski 2017; Swoyer 1991). For these reasons, Poldrack points us in the direction of the topographic organization of primary visual, somatosensory, and motor cortices as canonical examples of the sort of structural isomorphism that licenses the use of the term "representation".

There's a great deal to unpack here. We might start by noting that isomorphisms can be misleading. Primary motor cortex was for decades thought to have a somatotopic structure, such that it might, on this reading, be said to represent the body's effector organization. But contemporary work has shown that the apparent somatotopy is only approximate (Schieber 2001). For instance, adjacent body parts are often spatially intermixed, so that although there is an overall somatotopic trend, it holds strictly only for the high-level body plan. More strikingly, microstimulation of local regions of the cortex evoke complex movements involving multiple effectors and muscle systems (Graziano et al. 2002a). The apparent somatotopy was in fact a side-effect of the most efficient organization of a whole-body control system (Schieber 2001; Graziano et al., 2002b; Graziano 2011). While it is of course true that there is a systematic empirical relationship between local activity in M1 and bodily movements, that relationship would appear to be more naturally interpreted as a causal-functional one. Structural isomorphisms don't always license representational glosses.

A focus on structural isomorphism is also going to potentially rule out a number of neural systems that one might wish to include. Another putative example Poldrack offers of a representation in good standing is the activity of hippocampal place cells (O'Keefe & Dostrovsky 1971) which are said to represent particular locations in an animal's environment. As Poldrack notes, in this case we are *missing* the structural/spatial isomorphism that we see in the other

examples. Indeed, the most natural interpretation of hippocampal place cell activation is as an indicator or detector.[1] This would appear to threaten its status as representational. A related case is activity in the olfactory bulb, which has an extremely complex relationship to odor stimuli; it would be difficult to even identify straightforward correspondences between activity and odor (Freeman 1988). In sum, although it is widely acknowledged that an entity must meet criteria beyond having "a systematic empirical relationship" to a thing to represent it, the initial candidate suggested by Poldrack, structural isomorphism, appears neither sufficient (isomorphisms don't always indicate representation) nor necessary (representations needn't be strictly isomorphic) to earning the label "representation" in neuroscience. And here Poldrack faces the first of some important dilemmas raised by his account: he uses structural isomorphism to ground his case that NRs are legitimate representations, but insisting on structural isomorphism as a necessary feature of NRs would rule out some (perhaps many) of the things that neuroscientists in fact *call* representations. He is caught, that is, between developing a defensible account of NRs, and being faithful to the vocabulary of neuroscience.

Thus, although it is certainly true that neuroscientists are "comfortable" (Poldrack 2020:5) using the term "representation" to refer to the full range of examples canvased by Poldrack, it is equally clear that this usage does not—indeed, probably cannot—by itself establish the relevant legitimacy of this use, and "structural isomorphism" cannot do the necessary work, either. So how might we establish that NRs are indeed representations in good standing? Poldrack suggests that one way is to examine a class of systems—ANNs—where there are *provably* representations, because the system is designed to generate them, to argue that these artificial representations (ARs) are relevantly similar to NRs, and to establish that ARs are representations in good standing. By transitivity, this would support the claim that NRs are representations in good standing. In the next sections, we turn to this argument.

3. Representation learning in ANNs

As Poldrack notes, modern deep learning systems—he primarily discusses Hierarchical Convolutional Neural Nets—are the descendants of connectionist models and work according

[1] And perhaps not even this. For there is a larger issue raised by the assertion that activity in hippocampal place cells represents locations in the environment, one that is obliquely raised by Poldrack's discussion of receptive fields: hippocampal place cells are *not* in fact selective for places in any straightforward way. For instance, they fire not just when an animal is at a location, but also before an animal enters and after it leaves the location, with the difference being a matter of the relative timing of the activity with respect to background theta oscillations (Buckner 2010); this property might play a role in supporting spatial navigation (Stachenfeld, Botvinick & Gershman 2017). Moreover, they have been long known to often show odor selectivity, or a mixture of place and odor responsivity (Wood et al. 1999). The apparent straightforward correspondence between cells and spatial locations is in fact context dependent, and the activity of these cells is only interpretable given a known and constrained setting (Brette 2019). Here we merely flag this issue, but we will return to the point that the constraints imposed by environmental context go unrecognized here, and weaken Poldrack's case overall.

to some of the same principles, with some important twists. The basic idea is to have a set of input nodes to which one feeds the data to be processed, some number of layers of "hidden" nodes that do the processing, and a set of output nodes from which the results of the processing can be read. In the example described here, the input is an image, and the output is a label for the content of that image. Such systems are *extremely* successful at a number of useful tasks, including image classification, and the usual account given for their success is that they are able to learn (or "extract") the relevant features in the input data needed for the task. These features are the representations in representation learning. Poldrack offers us some examples of the features relevant to image classification in his Figure 3, panel a. We reprint a version of that figure from the original source.

Figure 1: Features extracted by an image classification HCNN. Reprinted from Olah et al 2018, licensed under CC-BY.

These are some of the feature representations that the HCNN learned during training, and that aid it in making classification decisions. As Poldrack notes, these are, at root, useful mathematical abstractions "ultimately defined by the numeric values of the parameters in the network." And yet, he writes, "[i]t is nonetheless difficult to resist the natural tendency to view these putative representations as having interpretable semantics falling at levels intermediate between the raw input and the final object category decision." (2020: 8-9) He offers such interpretations as "floppy ears" and "furry legs".

Here it is instructive to recall that, like the word "representation", "feature" has a long history in philosophy and psychology, and just as with "representation" it is far from clear that "feature" is more than a homonym when uttered by a psychologist and by a computer scientist. In the mouth of Anne Triesman, perhaps the most prominent of the psychologists investigating feature-based object perception, a "feature" refers to one of a number of identifiable elements into which an object might be decomposed (Triesman 1986; 1988). Examples she investigated included verticality, curvature, slant, color, intersection and the like. To be sure, some of her discoveries of basic features were surprisingly abstract, such as closure[2], and raise interesting questions of exactly what it means for an object to be "composed" of such features.

[2] The difference between a C and a O, both of which possess "curvature", is "closure".

Nevertheless, they are all (and possibly necessarily) introspectively available elements of our perception of objects. The features on display in Figure 1 above are nothing like this; those images are neither captured by, nor do they capture the meaning of phrases like "floppy ears". Even assuming that "floppy ears" is a good candidate for a feature of the human beagle representation, the HCNN that generated the features in Fig. 1 has not in fact extracted it.

Let us be clear on a point we are *not* making here: One of the reasons the HCNN-extracted features work so well, and there is a deep lesson here, is that they *escape* the naïve decomposition of the image that human scientists have long relied on to design experiments and build classification systems. The early classification system Pandemonium, for instance, relied on a naïve decomposition of letters into features like "vertical bar" and "slanted line" (Selfridge and Neisser 1960). It worked passably well for a narrow range of stimuli, but was extraordinarily brittle. Similarly for other early visual discrimination systems rooted in our common-sense analysis of the elements of our experience. We'd wager that this is a necessary feature: *no* system that used the analyzable features of our experience as its elemental units could possibly work.

And this gets us to the point we *are* making: the features on display in Fig 1 may be ARs, (and they may even be relevantly similar to NRs; more on this, below), but they are *not* structural representations. They do not bear a recognizable isomorphism to the images from which they were developed, and they are not semantically interpretable (Poldrack's reported temptation notwithstanding). Indeed, they are not in any straightforward way *features* of the images at all in that they cannot be described as being a part or element of that image the way curvature and closure are elements—features, properties—of a circle[3]. There is in fact nothing in the physical world to which these machine learning features correspond (although there are, of course, things in *mathematical reality* to which they correspond).

Now, and this is important, these features *do* bear a "systematic empirical relationship" to the images; and they *do* comprise images in the sense that a given feature vector can be used to *reproduce* an image. They have an absolute claim to being a crucial part of the mechanism for image recognition. And this brings us to the next of the dilemmas facing Poldrack, here. The more he wants to hold up ARs as the paradigmatic case of representation, the less claim he has that they illuminate the sorts of things that neuroscientists have in fact identified as representations such as tonotopic and somatotopic "maps", which are semantically interpretable and isomorphic to their referents, or even hippocampal place cell activation which is at least semantically interpretable.[4] But the more he hopes to hold on to these

[3] And given the importance of convolution to HCCNs there may be principled reasons to deny that this kind of compositional relationship is even possible.

[4] Again, under known, constrained conditions

common-sense characteristics of representations, the less claim there is that ARs are representations in good standing at all.

The problem is deepened by his clever argument that AR-like features can indeed be found in the brain, as illustrated in his Figure 3, panel b. To remind the reader, these features were generated by a deep-learning image synthesis process, and found to maximally stimulate neurons in primate temporal lobe. If that's what *actual* NRs look like, this opens a gap between them and the use of the term in neuroscience—and, not incidentally, between neuroscience and psychology. One as yet unremarked feature of hippocampal place cells and retinotopic maps is that they retain an intuitive and obvious relationship to—they can usefully illuminate—experience and observable behavior. The more the discoveries of neuroscience lead us towards endorsing AR-like features as the drivers of neural mechanisms, the less use they will be in illuminating experience and behavior. If *those* are the sorts of features that underly our perception of dogs, it's pretty odd that dogs look like dogs. But they exactly do.

Thus, granting that Poldrack has correctly described the linguistic practices of neuroscience, the account he develops of actual NRs—the features in his Fig. 3—appears to *exclude* things neuroscientists typically refer to, and furthermore undermines the main attraction of representational explanations in psychology, that they offer explanatory entities that can be related to elements of our experience. This does *not* mean that Poldrack is wrong that the brain trucks in entities of the sort he describes (be they representations or no); it was always a bad bet to expect the mechanisms of the brain to be isomorphic to our analytic breakdown of experience (Dennett 1989; Uttal 1979). But if we grant that NRs are in fact relevantly similar to ARs (and therefore have equal claim to being representations in good standing), this may suggest that neuroscientists are generally searching for the wrong thing when they go in search of representations. It may also mean that the current relationship between psychology and neuroscience (especially in cognitive neuroscience) needs to be rethought.

4. Do ARs meet the job description challenge?

Let us assume for the sake of this argument that actual NRs are in fact like ARs. Do they meet the job description challenge? Poldrack structures his argument that AR/NRs *do* meet the job description challenge around Nicholas Shea's criteria for representational explanation (2013, 503).

A representational explanation requires:

> (a) An explanandum concerning how the system operates or behaves in relation to its environment.

(b) A putative explanation of (a) that relies in part on attributing representational properties to the system (e.g., keeping track of p, aiming at q, etc.).

(c) An account of how the explanation in (b) succeeds (remaining open to there being no such account).

(d) If there is a positive answer to (c), a characterization of the kind of properties the representational properties of the system would have to be for the explanation in (b) to succeed in explaining (a) in accordance with the account (c).

He uses these criteria to give an account of the work of the primate inferior temporal cortex and HCNNs in object detection as follows:

(a) **The Object Recognition Question**: How does the system perform object recognition on input images? The explanandum is the system's ability to attribute a name to a specific object in an input image (2020: 12)

(b) **The Feature Detection Answer**: The system does representational work by decomposing the visual input into a hierarchy of features that enable object recognition. (2020: 12)

(c) **Architectural Suitability:** Feature detection is successful since the hierarchical convolutional architecture of the system reflects the "compositional and hierarchical structure of the macroscopic world that gives rise to visual images." (2020: 12) More generally, Poldrack asserts the:

(d) **Inductive Bias Claim**: Artificial neural networks such as HCNNs are successful at complex problems because a set of inductive biases "are built into the architecture, which are well-matched to the underlying function that relates visual images of objects to their category membership." (2020: 10)

Poldrack argues that criterion (d) in particular "encapsulates" Ramsey's "job description" challenge (2020: 12-13). He suggests the properties of the representations that enable object detection are required to meet the:

I. **Reliable Stimulation Condition**: The representation should be causally triggered by the presence of the stimuli. This is a necessary but not sufficient condition of the criterion (2020: 13).

II. **Structural Isomorphism Condition**: The structure of the content of the representations should match the "hierarchical and

compositional structure of the visual world" (2020: 13). He argues this condition is satisfied because of the:

III. **Dimensionality Reduction Claim:** The representations do the mathematical work of reducing the size of the problem space, from a high-dimensional manifold to a low-dimensional one.

IV. **Decouplability Condition**: The representations should be decouplable from the triggering stimuli, because it is difficult to explain the results of ontogenetic or (non-environmental) stimulation otherwise (2020: 13).

We will stipulate that that the AR/NRs detailed by Poldrack meet conditions I and III.[5] However, they fail to meet the job description challenge because they fail to meet conditions II and IV. We've already discussed our issues with Condition II. The features extracted from the input images are *not* imprints of structures of the physical world, nor are they straightforward decompositions of those objects into parts with preserved spatial relationships, etc. Nor has it been in any way established that that there is some higher-order isomorphism between the hierarchy of object features and the hierarchy of features extracted by the HCNN, sufficient to establish a representational relationship between these domains (Roskies 2021).[6] There is, of course, a mathematical mapping between the features and the original stimulus, but it would require a separate argument to show that such a mapping does the work that "isomorphism" is intended to do for representations. There is, after all, an available mathematical mapping between *any* two things. The question is whether *this* mapping has the characteristics required to support the representational claim. We doubt it. (We *don't* doubt this would be a deeply interesting discussion, but we will not pursue it further here.)

As for decouplability, Poldrack argues that the behavioral responses triggered by optogenetic technologies (which directly stimulate certain sets of neurons) are similar enough to responses

[5] Claim III may be problematic for some of the reasons we raise in section 5, but following out that intuition is a topic for another paper. (Literally: Champion, 2021).

[6] An anonymous reviewer suggests that it's just such a higher-level isomorphism that Poldrack is in fact arguing for, and on which his claims that NRs are representations in good standing depend. To be sure, one *can* use the similarity of representation spaces to argue that one thing, say Infero-Temporal cortex (IT), potentially or provisionally represents another, say object categories. This is the strategy followed in Mur et al. (2013) (see also Kriegeskorte et al. 2008) to argue that IT represents object categories, because the representational space generated by IT while participants were shown a variety of objects was highly correlated with the categorical representational space generated by analyzing similarity judgments made between the same objects, But this isn't the strategy Poldrack follows, here. Instead, he demonstrates the similarity between the hierarchies generated by primate cortex and the HCNN in the service of establishing that the intermediate representations—the ARs and NRs—are relevantly similar: "Thus, the intermediate representations learned by an HCNN…are likely to be very similar to the representations present in biological neurons…" (Poldrack 2020: 11). That is, Poldrack's explicit argument for the relevant similarity between ARs and NRs rests on comparing the *feature* representations; the similarity of the representational spaces helps establish this claim. It is this argument, that ARs (and NRs) are similar abstract feature representations, that we are critically evaluating here. This being said, all three anonymous reviewers noted that Poldrack's arguments can be a bit difficult to pin down—so we may certainly have gotten them wrong. Still, even a misunderstanding can promote eventual understanding if openly discussed.

elicited in the presence of real, present objects that these behaviors would be difficult to explain without offering a representational explanation. Clearly, in such cases a behavior is being performed in the absence of the relevant environmental target. True enough. But we'd like to dub this the "why are you hitting yourself?" account of decouplability, after the game that all older brothers played with their younger siblings. The sort of neural intervention in the mechanical order that Poldrack describes is of course more subtle than that, or at least requires more expensive technology, but it has that same flavor: a brute force causal intervention that compels a behavioral outcome. What isn't clear is why a representational explanation is called for. "Why are you hitting yourself?" has a perfectly straightforward answer: because you are making me! One needn't posit the errant representation of a mosquito being slapped.

There's a broader issue with this sort decoupling that is perhaps worth bringing to the fore, in part for its own sake, and in part because it reinforces one of the themes of this commentary. The *letter* of the decouplability requirement is simply that a representation can be triggered/tokened in the absence of its proper cause. But the *spirit* of the requirement showcases the *autonomy* of the agent: if I so choose, I can just think of my grandmother, though she has been dead for many years now. As Sarah Robins (2018) points out, one can use optogenetic intervention to cause mice to "recall" a location, but arguably *real* recollection is a mental act of retrieval performed by an animal for a behavioral purpose. Only the latter appears to naturally invite representational explanation. Two things follow: first, the letter-of-the-law formulation of decouplability may be too broad; second, and in keeping with the theme here, brute force interventions of this sort represent a neuroscientific, and not a psychological, version of decouplability.

5. Are representations necessary and universal?

Poldrack ultimately seeks to establish that representations are *necessary* for intelligent behavior. As we mentioned at the outset, this is an extremely bold claim that is not supported by his treatment of representations in neuroscience, ANNs, and his response to the "job description challenge."

1) Our first point along these lines is simple and straightforward: Representations in good standing cannot be universal and necessary to successful cognitive outcomes like classification, because ANNs use ARs, and ARs are not representations in good standing.

2) Furthermore, whether ARs are necessary and universal even in deep learning contexts is itself an open question. In particular, Hasson et al. argue that overparameterized "direct-fit" deep learning models make effective predictions while remaining "effectively agnostic to the underlying [generative] structure of the world" (2020: 423). Whereas Poldrack takes representation of a generative process as necessary for effective generalization (15), overparameterized "direct-fit" models make effective predictions by interpolation, rather than

extrapolation, on big data (418). Thus, the question of whether representation learning is necessary for deep learning requires further investigation.

3) The Dimensionality Reduction Claim undertakes too much. While ARs do help with dimensionality reduction, they are not the source of *physical* constraints producing "natural images" (13). Poldrack acknowledges the role of physical constraints elsewhere (10) but fails to pay them any substantial dues in his account of object detection via ARs. ARs are not needed to reduce the dimensionality of the space of *all* possible pixel values, since environmental regularities already significantly reduce the size of the problem space.

4) For all the reasons canvassed here, we don't think that the success of "representation learning" implies that RTM has been vindicated, as it might be thought to, especially if it had been established that representations were necessary and universal. Even if AR-like entities turn out to be necessary and universal, these are apparently very different from the entities posited by RTM. The term "representation" has rather different meanings in its different contexts, and inferences from one domain—computation—to the other—mind—are severely curtailed. Nevertheless, as new technologies and techniques from computer science continue to be developed, it is certain that they will continue to have an impact on how we understand and study the brain and mind.

6. Concluding remarks

Here we hope to have shown that Poldrack has indeed accurately characterized the scope of the term "representation" as it is deployed in the neurosciences. However, in endeavoring to solidly establish the existence of such entities, he identifies a narrower class of things, NRs, that neither answer to the term as used, nor meet the "job description challenge". This would appear to lead to the conclusion that NRs are *not* MRs, nor in any straightforward way related to them (beyond what one achieves by hand-waving about causal underpinnings). Contrast this outcome to the hope behind the empirical relationships identified early in the paper: this neural activity means "I am here", and this "hearing a 120Hz tone", or "seeing an edge". Actual NRs as identified just don't seem to be interpretable in terms of experience or knowledge, and this appears to open up a gap between neuroscientific and psychological explanation. This was probably not Poldrack's intention, but we find it to be an interesting consequence of his project, worth noticing.

Still, if we are going to rekindle the "representation wars" (Clark 2015) we should appreciate the context that gave rise to them and note how different the current situation in fact is. It is true on anyone's theory of the brain (or the brain-body complex, or whatever the "cognitive" system turns out to include) that it is a machine for turning perception into adaptive – correct by some measure – behaviors, which behaviors also, for living organisms, and perhaps also for some artificial agents, involve generating further perceptions. What is at issue is the nature of the mechanism that modulates this reciprocal perception-action coupling. Where

representations came into the picture, canonically in Fodor's work on the Language of Thought Hypothesis and RTM (Fodor 1975; 1981), was with the idea that rationality needed special explanation, and that explanation was the existence of representations and predicates and rules and the rest of the cumbersome baggage of cognitivism. But what Dennett (1981) shows us is that rationality isn't super-added to adaptive behavior; it is its necessary condition. In this sense, you get rationality for free, and its presence provides no bias toward any potential answer to the question of mechanism.

Putting our bets on the table, we suspect that few of the elements of the mechanisms responsible for adaptive behavior will be best understood as MRs (of the sort posited by RTM), for they will act as both information-carrying *and* control structures. Such structures needn't be isomorphic to the world in the way that remains central to accounts of "representations in good standing", and to Poldrack's account of representations (although not NRs) here. What they need to be, and what Poldrack's NRs are, are information-carrying patterns of neural activity that play a role in the control of adaptive behavior. Poldrack, in his way, recognizes this when he adopts the apparently ecumenical position that dynamic systems explanations may well be what's needed for motor control, but that, nevertheless, representations and representational explanations are indispensable to perception. In fact, this position simply accepts and reinforces the perception/action dichotomy at the center of RTM, which more ecologically-oriented proposals reject.

In the end, we rather like this account of NRs, because even a Gibsonian neuroscience can take them fully on board (Raja & Anderson 2019). These "patterns of activity that bear a systematic relationship to the structure of the external world and play a causal role in behavior" (Poldrack 2020: 16) simply have no implications for the shape of our psychology; since NRs do not imply MRs, we are free to follow ecological principles of looking for multi-level constraints and law-like relationships between environments, organisms and behavior, knowing that the mechanisms making adaptive behavior possible are more surprising than we had anticipated.

References

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*, e215. https://doi.org/10.1017/S0140525X19000049

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, *61*, 27-48.

Champion, H. (2021). Representation without constraint: a critique of Poldrack's account of object detection using artificial neural networks. *Rotman Graduate Student Conference*, Western University, May 2021.

Clark, A. (2015). *Predicting peace: The end of the representation wars*. Open MIND. Frankfurt am Main: MIND Group.

Dennett, D. C. (1989). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1981). "Intentional systems" In: *Brainstorms: Philosophical Essays on Mind and Psychology* (pp 3-22). Cambridge, MA: MIT Press.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1981). *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Freeman, W. J. (1988). Nonlinear neural dynamics in olfaction as a model for cognition. In: T. Melnechuk, ed., *Dynamics of Sensory and Cognitive Processing by the Brain* (pp. 19-29). Berlin: Springer.

Gallistel, C.R. & King, A. (2009). *Memory and the Computational Brain*. Malden: Wiley-Blackwell.

Gładziejewski, P. & Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology & Philosophy*, *32***,** 337–355. https://doi.org/10.1007/s10539-017-9562-6

Graziano, M. S. (2011). Cables vs. networks: old and new views on the function of motor cortex. *The Journal of Physiology*, *589*(Pt 10), 2439.

Graziano, M. S., Taylor, C. S. & Moore, T. (2002a). Complex movements evoked by microstimulation of precentral cortex. *Neuron*, *34*(5), 841-851.

Graziano, M. S., Taylor, C. S., Moore, T. & Cooke, D. F. (2002b). The cortical control of movement revisited. *Neuron*, *36*(3), 349-362.

Hasson, U., Nastase, S. & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron* 105(3), 416–34. https://doi.org/10.1016/j.neuron.2019.12.002.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K. & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A. & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128. https://doi.org/10.3389/fpsyg.2013.00128

Nirshberg, G. & Shapiro, L. (2021). Structural and indicator representations: a difference in degree, not kind. *Synthese*, *198*(8), 7647-7664. https://doi.org/10.1007/s11229-020-02537-y

O'Keefe, J. & Dostrovsky, J. (1971). The hippocampus as a spatial map preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3*(3), e10. https://distill.pub/2018/building-blocks/

Poldrack, R.A. (2020). The physics of representations. *Synthese.* https://doi.org/10.1007/s11229-020-02793-y

Putnam, H. (1967). "Psychophysical Predicates" In: W. Capitan and D. Merrill (eds), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press.

Raja, V. & Anderson, M. L. (2019). Radical embodied cognitive neuroscience. *Ecological Psychology*, *31*(3), 166-181.

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Robins, S. K. (2018). Memory and optogenetic intervention: separating the engram from the ecphory. *Philosophy of Science*, *85*(5), 1078-1089.

Roskies, A. L. (2021). Representational similarity analysis in neuroimaging: proxy vehicles and provisional representations. *Synthese*. https://doi.org/10.1007/s11229-021-03052-4

Rupert, R. (2018). Representation and mental representation. *Philosophical Explorations, 21*(2), 204–225.

Schieber, M. H. (2001). Constraints on somatotopic organization in the primary motor cortex. *Journal of Neurophysiology*, *86*(5), 2125-2143.

Selfridge, O. G. & Neisser, U. (1960). Pattern recognition by machine. *Scientific American*, *203*(2), 60-69.

Shea, N. (2013). Naturalising representational content. *Philosophy Compass* 8(5), 496–509. https://doi.org/10.1111/phc3.12033.

Stachenfeld, K., Botvinick, M. & Gershman, S. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20***,** 1643–1653. https://doi.org/10.1038/nn.4650

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87*(3), 449-508. http://www.jstor.org/stable/20116918

Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, *255*(5), 114B-125.

Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology Section A*, *40*(2), 201-237.

Uttal, W. (1979). Do central nonlinearities exist? *Behavioral and Brain Sciences*, *2*(2), 286.

Wood, E.R., Dudchenko, P.A. & Eichenbaum, H. (1999). The global record of memory in hippocampal neuronal activity. *Nature*, *397* (6720), 613–616.