Forthcoming in Synthese: Special Issue on Scientific Realism in Cognitive Neuroscience
Eds. Inês Hipólito, Thomas van Es, Guilherme Sanches Oliveira, Matteo Colombo

What is a Theory of Neural Representation For?

Andrew Richmond

Abstract

This paper asks how representational notions figure into cognitive science, especially neuroscience. Philosophers have a way of skipping over that question and going straight to another: *what is neural representation?* What is the property or relation that representational notions pick out? I argue that this is a mistake. Our ultimate questions, as philosophers of cognitive science, are about the function and epistemology of cognitive scientific explanations — in this case, explanations that use representational notions. To answer those questions we must understand what representational notions contribute to science: what they enable scientists to do or explain, and how. But I show that we can do this without raising traditional and vexing questions about the definition of neural representation, or the nature of a property or relation that notion picks out. Taking this approach, I defend a realist account of representational explanation that underwrites important connections between philosophy and neuroscience.

1 Introduction

Representational notions seem essential to our understanding of the brain. Neuroscience tells us that the brain supports navigation by representing spatial properties (Behrens et al. 2018), classifies objects by representing their features (Chang and Tsao 2017), supports language use by representing word meanings (Borghesani and Piazza 2017), and so on. So a central question in philosophy of neuroscience has become, *what is neural representation?* What is it for some neural structure or activity to be a representation, and to represent what it represents?

Though this is a central question in philosophy of neuroscience, it is not a *fundamental* one. That is, the question does not hold its central position because of its intrinsic interest, but because of its relationship to deeper questions. It is because we want to understand how and why neuroscientific explanations work that we worry about the properties they might refer to and how to define those properties. The main contribution of this paper will be to argue for a way of understanding representational explanation that does not detour through debates about the nature or definition of representation. Instead, I will discuss what representational *notions themselves* contribute to neuroscience, emphasizing how they help neuroscientists construct and understand models of the brain's causal structure.

In Section 2 I'll prepare the way with some examples of representational explanation and the questions they raise. In Section 3 I'll give the account of representational explanation that I alluded to a moment ago. In Section 4 I'll connect this account to debates about scientific realism. And in Section 5 I'll conclude by discussing some upshots and objections.

2 Representational explanations, and the questions they raise

I'll start with two brief examples of representational explanation. First, consider explanations of navigation. Many organisms have a remarkable capacity to navigate their environments, avoid obstacles, find remembered locations, and travel home from new places along efficient paths. Our current understanding of how the mammalian brain supports navigation started to come together in the '70s with the discovery of place cells: the brain's spatial representation system. Cognitive scientists had long suspected that the brain navigated using a neural map of its environment (Tolman 1948), and place cells seem to be a part of that map. They respond selectively to locations in the environment, and together with grid cells that tile the environment, they seem to combine information about the distances an animal has traveled in different directions (from neurons that represent distance and direction, e.g. by representing head direction) to represent the animal's current distance and direction from previous locations (Moser et al. 2017, p. 1451). In short, navigation appears to be possible because the hippocampus maintains a coordinate system supported by path integration algorithms that derive representations of an animal's location from representations of its previous movement directions and distances.

Second, consider our capacity to recognize and distinguish between faces (Kanwisher and Yovel 2006). On the traditional view, this ability is supported by neurons in the fusiform face area (FFA) that respond selectively to faces. Those neurons appear to derive representations of objects as faces from representations of face-parts (eyes, mouth, nose), the spatial layout of those parts, and the bounding contour typical of faces (Kanwisher and Yovel 2006). They also appear to *individuate* faces — to represent faces as the particular faces they are — because their activity is largely invariant across different presentations of the same face, though this invariance is imperfect in important ways (Kanwisher and Yovel 2006). There is debate over how FFA individuates faces, but an interesting suggestion is that it does so by representing the precise way that different faces deviate from a "norm or average face" (Kanwisher and Yovel 2006).

These are controversial areas of research. E.g., there is debate over FFA's selectivity for things other than faces and its potential role in capacities other than face-perception (Rhodes et al. 2004), and a corresponding debate about whether it represents faces specifically or a broader category (Kasper et al. 2022). Similarly, there has been debate about whether place cells represent an animal's *current location* or its *intended direction of movement* (Euston and McNaughton 2006). I don't intend to intervene on those debates here (but I'll return to both later). The examples are just intended to illustrate the role that representational notions play in neuroscientific explanation, to help me illustrate the questions they raise.

Those questions differ between neuroscience and philosophy. Neuroscientists, of course, ask all kinds of questions about representations: where are they; what are they doing; what capacities do they support; what do they represent? (Or do they represent at all?) I'll return to these questions in Section 5. Philosophers tackle more fundamental questions about representational explanation *qua* mode of explanation. What is the function and epistemic status of representational explanations?

What are these explanations doing for cognitive science? How do they work? And *why* do they work — why are they successful, if and when they are?

Since these are questions about representational explanation *qua* mode of explanation, one might expect them to hinge on views of explanation itself.[1] But most work in this area is non-committal about the deeper nature of explanation,[2] and there is good reason for this. What matters about representation, and what it may be helpful for philosophers to clarify, is how it contributes to the various scientific activities that advance our understanding of the brain: modeling, theorizing, predicting, and so on (see Waters 2019, on the variety of scientific activities). And how we explain that contribution will probably not depend in any significant way on whether we accept (e.g.) an ontic or semantic view of explanation (Shea 2018, p. 88, fn. 16). Perhaps more concrete views of scientific explanation, e.g., mechanistic (Craver 2007) and non-causal (Chirimuuta 2018) views, are more relevant? It would be possible to argue for, say, the mechanistic view, and then allow it to constrain one's account of representational explanation. But I think it will be more fruitful to understand representational explanation on its own terms, letting the conclusions about mechanism (and other views) fall where they may. So I will follow the rest of the literature by using a broad and unconstrained understanding of explanation, aiming to capture the role of representational notions in the various things neuroscience does to investigate and understand the brain.[3]

With that in mind, we can make the previous point a bit more clearly. Philosophers have been interested in how and why representational explanation works, that is: what it contributes to the scientific understanding of the brain. The *standard approach* in philosophy holds that this question is best answered by investigating the nature or metaphysics of representation, via some kind of definition. This is perhaps best understood as a simple, and quite sensible, three-step tactic:

[1] I'm grateful to two anonymous reviewers for pressing me to clarify this.

[2] Cummins' well-known book on representation (1991) is a striking example: it makes almost no mention of his earlier and very influential book on psychological explanation (1975), even though the later book is explicitly concerned with the explanatory role of representation in psychology (1991, p. 2).

[3] My account will stress the psychological role of representational notions, so I'll be committed to the idea that there is a psychological dimension to explanation. I think this will be uncontroversial, though, given that "explanation" is understood as the broad set of scientific activities that advance our understanding of the brain.

Step 1 Note the ubiquity, and perhaps success, of a distinctive type of explanation — in this case: explanations that use the notion of representation. Given their ubiquity and perhaps success, it is important to understand *how and why these explanations work*.

Step 2 Give a plausible skeleton answer to those questions: the explanations work by attributing a special property or status to part of the brain — REPRESENTATION, or REPRESENTATION OF X (faces, locations, ...).

Step 3 Flesh out this answer by saying what exactly that property *is*. What is it for some part of the brain to be a representation (or a representation of *x*)?[4]

Using this tactic we can move quickly from difficult and nebulous questions about how and why representational explanations work to specific and tractable questions about *what representation is*. This is not to abandon questions about representational explanation for metaphysical ones; it is to *pursue those questions through* metaphysical ones. If representational explanation works by attributing a special property or status to certain activities/structures in the brain, then to understand representational explanation we must understand what the activities/structures contribute to an explanation *in virtue of having* this special property. And that means understanding what it is to have the property in the first place. That would not just help answer the philosopher's questions, but the neuroscientist's too: a neural structure or a bit of neural activity will either meet the criteria to be a representation, and a representation of *x* for any given *x*, or not. That will, in principle, tell us which parts of the brain represent and what they represent.

This approach has generated illuminating work. Cummins' foundational book on mental representation is one example. He writes:

> Empirical theories of cognition can and do take the notion of mental content as an explanatory primitive. But this is a kind of explanatory loan. . . . If it turns out that

[4] I'll mostly suppress the parenthetical ("or a representation of *x*") from now on. The standard approach applies whether we are focused on the property of being a representation *simpliciter*, or of being a representation *with particular content*. We'll see examples of both in the coming pages.

> the notion of mental representation cannot be given a satisfactory explication — *if in particular, no account of the nature of the (mental) representation relation can be given that is consistent with the empirical theory that assumes it* — then, at least in this respect, that empirical theory must be regarded as ill founded. (Cummins 1991, p. 2, emphasis mine)

Step 2 is so natural that Cummins can glide over it and move straight from Step 1, a recognition that theories of cognition use the notion of representation or content (the first sentence), to Step 3, the question of what precisely representation is (the second and third sentence). Given a different assumption at Step 2, this transition would be a non-sequitur. It is only because we assume the explanations work by attributing a special property or relation to the brain that we need an account of what that property or relation is.

This assumption is dominant in recent work as well. E.g., Shea begins his account of representational explanation by moving, just like Cummins, straight from the existence of representational explanations to puzzles about what exactly representation is, like the following: "That mental representations are about things in the world, although utterly commonplace, is deeply puzzling. How do they get their *aboutness*? The physical and biological sciences offer no model of how naturalistically respectable properties could be like that" (Shea 2018, p. 5).[5] More examples are not hard to find. Even Ramsey, who is especially explicit that his primary concern is with representational *explanation*, still aims to understand it by analyzing "the sort of physical conditions and relations that have been assumed to bestow upon an internal state the *status* of representation" (Ramsey 2007, p. 189, emphasis mine). And he proposes his own set of conditions: to be a representation is to have a particular functional role or satisfy a certain "job description" (Ramsey 2007, p. 24).

This focus on neural activity instantiating some property, satisfying some definition, or meeting some criteria to count as a representation is the defining characteristic of the standard approach,

[5] Both Cummins and Shea are after a *naturalistic* metaphysics of representation, but that isn't the part of their views I'm targeting. Even if they went on to argue that representation is a certain orientation of Cartesian substance, they would still be pursuing the standard approach as I've described it, through the three-step tactic.

and it is what my approach will abandon. I'll take a different tack at Step 2 of the three-step tactic, which will call for a different approach to Step 3 as well:

Step 1 Note the ubiquity, and perhaps success, of a distinctive type of explanation — in this case: explanations that use the notion of representation. Given their ubiquity and perhaps success, it is important to understand *how and why these explanations work*.

Step 2* Give a plausible skeleton answer to those questions: the explanations work by using representational notions to introduce conceptual resources that help serve neuroscience's explanatory goals.

Step 3* Flesh this out by saying what resources representational notions introduce, and how those resources help serve neuroscience's explanatory goals.

If you really wanted to miss the point, you could collapse this into the original three-step tactic: one thing a concept can do to serve neuroscience's explanatory goals is to refer to a property or relation: REPRESENTATION. But the point is that concepts can do other things for science, and investigating those other things is a promising way to understand the role of representational notions in neuroscientific explanation. The benefits of taking Step 2* will be clearest when we've seen the account of representational explanation it results in, so I'll turn to that momentarily.

First, though, note that the standard approach is present in neuroscience as well as philosophy. For the most part, neuroscientists take a pragmatic tack, using a workaday notion of representation and thinking not at all about its definition or metaphysics. A look at any cognitive neuroscience journal will show plenty of concern for *representations*, but no concern for the kind of issues that an account of *what it is to be* a representation would have to tackle, like whether one's account of representation includes things that are (arguably) not representations.[6] But sometimes neuroscientists do seem to be taking the same approach as philosophers. E.g., Eliasmith and

[6] This is often true even when scientists frame their questions in the metaphysical or definitional terms philosophers use. Palmer (1978) is an instructive example of this.

Anderson set out to understand the nature and significance of "representational claims," like the claim that some area of the brain represents some property (2003, p. 5). That looks a lot like Step 1, as I described it above (leaving aside any difference between representational "claims" and "explanations"). And they follow through on the three-step tactic like the philosophers I've cited, assuming that representational claims attribute a special relation, REPRESENTATION, to the brain, and that to understand representational claims we must "determine the exact nature of the representation relation" (Eliasmith and Anderson 2003, p. 5). So my targets are not just philosophers, but also the rare neuroscientists who think that a definition or metaphysics of representation is a prerequisite for understanding representational explanation (see also Baker et al. 2022; Poldrack 2020).

3 An account of representational explanation

My basic claim is that representational notions provide a way of imaginatively projecting the structure of one domain onto another: that's the resource they introduce to neuroscience. I'll flesh that out with some examples, building from simpler to more complex and relevant ones.

The simplest example concerns engineering. If you're arranging electrical circuits to build a computer, you're probably going to think of the circuits as composing gates that represent logical functions, and think of the inputs to and outputs from those gates as representing a pair of mathematical objects: 1s and 0s or Ts and Fs. What does this do for you? It helps you impose the abstract structure of the logical functions onto the causal structure of the gates, by connecting the gates so that their causal structure mirrors that abstract structure. In other words, the abstract structure acts as a model, and, by thinking of the gates as representing parts of that model, you cognitively link the two domains — wires and gates to mathematical and logical functions. These cognitive links can be seen most clearly in the way that, if you understand the parts of the circuit as representing parts of the model, you will come to talk about the circuits *in terms of* the model, and in terms of the model's domain. That is, you'll say things like, "if I put in a 1, I should get out a 0" or “the output of the AND-gate should be T in these conditions” — describing the system not in terms of its own (electrical and physical) properties, but in terms of the features or domain of the model you think of it as representing.[7]

What if, instead of engineering a computer, you were reverse engineering one? What if you found a computer on the beach somewhere and you wanted to understand how it worked? You would probably do the same thing, just without the freedom to alter the computer. After getting a rough impression of its input–output profile and its internal causal structure, you would propose hypotheses about the computer in terms of mathematical or logical entities you think of the inputs and outputs as representing. You would describe the input–output profile in terms of the ‘represented’ entities by hypothesizing that the computer adds numbers, computes logical functions, etc., outputting numbers, truth values, and so on, in response the same given as inputs. That is, you would describe its structure at a coarse grain in terms of a mathematical or logical model. And you would describe the internal, more fine-grained causal structure of the computer with algorithms that compute the relevant functions, e.g., describing structures as AND-gates or electrical impulses as 1s and 0s. In other words, you would describe the internal processes as well as the inputs and outputs as representing features of the model or the domain the model is defined over, and this would provide the same link between physical system and model that we saw in the forward-engineering example.

To summarize these examples, understanding the computer as representing logical entities means understanding it *in logical terms* in order to project the structure of a logical model onto the computer, as a model of its causal structure. This helps identify a space of models for the computer: ones with input–output structures that capture the behaviors we are interested in (like regularities in the computer’s production of outputs from inputs). It provides an intuitive and relevant way of describing the computer’s causal structure: in terms of the computational formalism that describes the model (cf. Richmond n.d.*a*). And it provides an intuitive link between the causally relevant

[7] To preempt an objection, this need not mean that the computer actually represents those logical operations and the entities they are defined over, or that its representing them should be our focus. In other words, what I’ve said so far is no motivation to revert from Step 2* to Step 2. You can imagine the problems of indeterminacy that would result if you took the fact that *x* is a model of *y* to constitute a representation relation between *x* and *y* (Sprevak 2010; Shagrir 2001).

parts of the computer and the aspects of the model they must correspond to or correlate with for the model to be accurate: the parts we think of them as representing.

But models need not be defined over abstract or mathematical objects. Compare an actual computer found on a beach (or close enough): the Antikythera mechanism. This ancient Greek device, commonly known as the first computer, calculated astronomical relationships. Since it was discovered, the Antikythera mechanism's inputs and outputs, as well as its internal causal structure, have been understood to represent astronomical entities and their relations, and have been modeled in terms of those entities and relations (e.g., Seiradakis and Edmunds 2018; Edmunds 2014). E.g., we see debates over models of a pin-and-slot device in the mechanism — whether to model it with *this* function or *that* one — cast as debates over what the device represents — *this* relationship or *that* one. In line with the previous examples, this representational thinking licenses descriptions of the pin-and-slot device *in terms of* the astronomical entities and relations that are used to model it — namely, the lunar inequality (Carman et al. 2012)[8] — linking the mechanism to the features of the cosmos that are being pressed into service as models of its causal structure.

So here, again, representational thinking helps us identify models based on the phenomena or behavior we are trying to explain (tracking astronomical relationships). It gives us a perspicuous way of describing our target system in terms of those models (describing the dynamics of the pin-and-slot device in terms of the lunar inequality). And it helps us clearly and explicitly relate our models to the causal structure of our target systems (we know what parts of the model, i.e., the cosmos, the pin-and-slot device's states must correspond to if the model is to be accurate). Does the Antikythera mechanism *really* represent the sun and the moon, in a philosophically rigorous sense? Maybe, but this has no bearing on my point. The point is to elucidate what representational notions allow us to *do* when we are trying to understand a complex system. They allow us to impose structures from the 'represented' domain onto the 'representing' system as models, not just when

[8] This is even more pronounced in popular treatments, where, e.g., a behavior of the mechanism that is modeled by relationships between the sun's motion and the moon's is described like so: "Put in the sun, get out the moon" (Marchant 2008, p. 144). Internal structures like gear trains are described similarly: "the motion of the sun [is] subtracted from its lunar equivalent" (Marchant 2008, p. 148).

we engineer a system but when we reverse engineer one — in the cases above, reverse engineering its causal structure insofar as that structure supports a capacity defined over some external domain (adding *numbers*, tracking *planets*). To return to the focus of this paper, you may have noticed that the previous sentence is nearly identical to a common description of cognitive science's main goal: to reverse engineer the brain by constructing models of its causal structure insofar as that structure supports cognitive capacities (Dennett 1994; Mekik and Galang 2022), where cognitive capacities are understood, like in the cases above, as abilities to produce environmentally-defined outputs (like numbers and planets locations, or movements and environmental manipulations) in response to environmental states of affairs.[9]

E.g., FFA, together with the visual processes leading up to it, is understood to take low-level environmental features as input and give categorizations of entities as faces (or particular faces) as output. Just as in the previous examples, there is an environmental structure or relation at play — not a relationship between the sun and the moon, but a relationship between *the low-level features of an object* and *its status as a face/non-face*. If the brain transitions from a registration of low-level environmental features to a reliable categorization — i.e., to a state that correlates with something's being a face/non-face — then the relationship between a stimulus' low-level features and its status as a face/non-face will be an accurate model of (one aspect of) the brain's causal structure at a coarse grain. And, just as in the previous examples, we can think of FFA as representing faces to project that relationship, that function from low-level to high-level features of the stimulus, onto the brain as a model of its causal structure at a coarse grain. This also identifies a space of finer-grained models: the structures that relate the low-level and high-level features, i.e. the processes that would compute or implement the function from low- to high-level features. It provides a perspicuous way of describing FFA: in terms of the model, i.e., in terms of the relevant environmental features and the relationships between them. And it links the model to the causal structure of the brain by highlighting the way that parts of the brain should correspond to features of the model, i.e. the

[9] Though this must include certain *internal* inputs, outputs, and states of affairs (like subjective experiences, goals, or modifications to the brain itself).

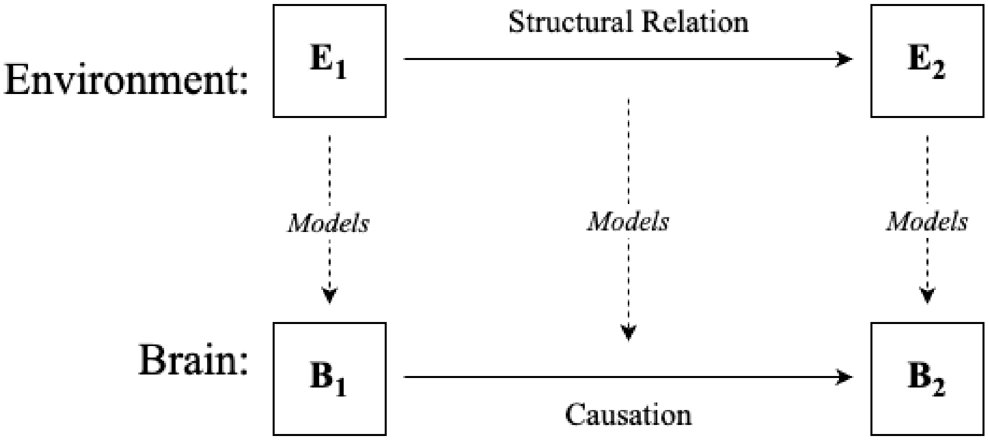

Figure 1: If the brain transitions from an internal state (B_1) corresponding to one environmental state (E_1) to an internal state (B_2) corresponding to another environmental state (E_2), the relation between E_1 and E_2 is a potential model of the brain at a coarse (or input–output) grain, and the structural relationship between E_1 and E_2 is a potential model of the causal structures in the brain that effect the transition from B_1 to B_2.

environment, if structures defined over those features are going to be good models of the brain. In short, representational notions do all the work here that they did in the examples above. They help us identify structures in the environment to use as models of both an organism's *behavior* and the causal structures that generate it (see Figure 1).[10]

FFA may be a simpler case than most (at least on the traditional view of FFA I've sketched) but the same story can be told elsewhere. In the case of navigation, we model the brain with a function from past features of the environment to an action or a future feature of the environment. E.g., consider a mouse that reliably finds the most efficient path home from a foraging trip. The relevant environmental structure is the relationship between the path the animal has travelled (particularly the directions and distances of its sub-paths) and the path back to its starting location. As I mentioned in Section 2, that path back is often close to the most efficient one available. Because it reliably moves from one set of paths to a particular further path — from the set of distances and directions travelled to the most efficient route back to its starting point — the mouse's behavior at a coarse grain will be accurately modeled by a function from the distances and directions travelled to the most efficient path back to its starting point. And its causal structure at a fine grain will be accurately modeled by

[10] Compare Egan (2014) and Cummins (1991) in relation to the figure, and see Section 5 for more on Egan's view.

some process that implements that function, like a path-integration algorithm. Here, again, the point is that by helping us project both coarse- and fine-grained environmental structures onto the brain, representational notions help us generate and understand *models* of the brain.

To summarize this account it is tempting to invert a popular slogan: the brain doesn't model the environment; the environment models the brain. Or rather, *we use* the environment to model the brain, and we do that by thinking of the brain in representational terms. Naturally, this kind of thinking comes in degrees. It is possible, on my account, to understand a Watt governor representationally — nothing is stopping you from thinking of its parts in terms of their environment and trying to use the structure in its environment to model it. But we generally understand the dynamics of a Watt governor without detouring through its environment, except perhaps to describe its inputs and outputs in terms of vehicle speed and combustion rate. Where those inputs and outputs — defining a high-level function that models the governor — are described in environmental terms, we can say that representational explanation is present in a very weak form.

But *thoroughly* representational explanations will model a system's *internal* structure with algorithms or processes whose stages or transitions are themselves defined over environmental entities.[11] E.g., we do not just model the brain as moving from a state corresponding to low-level visual features of an object to states that correspond to that object's status as a face/non-face. We model it as doing this via algorithms that are themselves defined over further environmental variables. On a cartoon version of this explanation: from sensory input the brain derives the locations and orientations of edges in a scene, from those edges the shapes, from those shapes the objects, from those objects the spatial relations between them, and from those objects and spatial relations the categorization of the object as a face or non-face. We describe the brain in terms of environmental relationships, not just between an object's low-level features and its belonging to the category face/non-face, but between those low-level features and many intermediate ones. And we use these environmental relationships to model not just the brain's input–output behavior, but the *steps* between input and output. That makes for a *thoroughly representational explanation*.

[11] The examples in Burnston (2020) may be borderline cases.

With all this in mind, we can say a few very general things about when representational explanation might be a good modeling strategy. First, it is essential that our interest in the target system is to explain how it brings about capacities defined over some domain outside the system[12] (whether an abstract or environmental domain) — e.g., the capacity to recognize faces, or to get from one place to another. Otherwise, external structures won't necessarily be accurate models of the behavior we're interested in, even at a coarse grain. Representational explanation will be most useful when the target system is also complex, necessitating some strategy for navigating a large space of possible models, and a strategy for clarifying and highlighting the models' explanatory connection to the target system's capacities (cf. Kriegeskorte and Diedrichsen 2019, pp. 408-409). And representational explanation is most likely to provide accurate models when the target system has evolved or been designed to get around with respect to certain environmental structures, and where dynamical short-cuts (simple transitions through state space that implement the input–output function) are unlikely. Design and evolutionary selection are often described as processes that impress the structure of the environment onto the systems undergoing selection (Shepard 1984). If the hippocampus has internalized causal structures recapitulating environmental structures, and can therefore be accurately modeled by those structures, this is at least partly because we faced selection pressure to navigate within those environmental structures.

Of course, design and evolution will not always impress the structure of the environment onto a system. The Watt governor was designed, and could no doubt have been selected for. And, more importantly, neither evolution nor selection appear necessary for representational explanation to apply accurately and fruitfully — that is, for us to be able to model one system using structures borrowed from another domain by describing the system in terms of those structures (cf. Richmond n.d.*b*). To put a finer point on it: evolution and design are not necessary or sufficient for the success of representational explanation, nor do they tell us how representational explanation works. They only provide *some* conditions under which representational explanation *may* be (but is not at all guaranteed to be) fruitful, given the non-teleological account of it I've described.

[12] Or outside the particular system component we're interested in (cf. Egan 1999).

The conclusion so far is that representational notions give us a way of identifying, and projecting onto the brain as models, environmentally-defined structures that, if they did accurately model the brain, would explain cognitive capacities. To return to the start of this section, representational explanation is a strategy that has all the benefits a logical model of a computer has over a description of it in electrical and physical terms. And as I've advertised, this is an account of what representational notions allow us to do, not an account of *what representation is*. To drive this home, consider how hopeless the account would be as the latter. Any system can be modeled by structures from a huge variety of domains. If the hippocampus represents every domain containing structures that could model it — even just domains with structures that could model it extremely well, and for non-trivial reasons — then everyone's hippocampus would represent everyone else's hippocampus, the mouse hippocampus would represent any computer programs we develop to do navigation in a similar way, and so on.

But instead of giving a metaphysics of representation, I've been describing the way representational notions help cognitive scientists understand the brain. And I've been able to do that without entering into traditional debates about what representation is — debates that even sophisticated and like-minded accounts, like Ramsey (2007), find themselves mired in. On my view, e.g., there is no question whether indicator representations *really are* representations Ramsey (2007, pp. 190-203); there is only the question of whether representational notions help us think about and model particular systems in the way I've described. This also distinguishes my account from isomorphism theories of representation (Cummins 1991; Ramsey 2007), as does the fact that isomorphism plays no role on my account except insofar as what makes *any* explanation involving a model appropriate is, partly, an (approximate) isomorphism between the model and its target system. What is distinctive about representational explanation is not the isomorphism that all model-involving explanations involve, but the specific way that representational notions help us construct and understand models.

To round off this section, I want to turn to a potential worry — one that will also let me illuminate a feature of the account. Take FFA again. The function from an object's low-level visual features to its being a face/non-face provides a good model of the brain only if the brain's causal structure actually mirrors that function. But we know it doesn't — not perfectly. 'Face' categorizations are sometimes given in response to non-faces, and vice versa. Prima facie, this should be a problem for my account. If the models generated by using representational notions aren't even accurate, how can they be explanatory?

Actually, though, the use of representational notions is an especially fruitful strategy when we are studying capacities that *do* fail a significant amount of the time, because it gives us resources to conceptualize those failures. Face-recognition has some illuminating patterns of error (e.g., pareidolia and prosopagnosia) that we might want a model to capture and explain. But we want a model that captures face-perception's successes and some of its more interesting failures, not a model that captures every failure due to noise, or a subject's boredom, distraction, tiredness, over-caffeination, etc. Including those failures would allow us to build a more detailed and accurate causal model of the brain, but this would not offer explanatory gains sufficient to justify the complexity of that model. Nor would the model connect as meaningfully to our explananda, which is not the whole pattern of face-categorizations we make, but the striking success of those categorizations. The cases of interest are the majority in which we *do* mirror the relationship between environmental input and an object's actual category, and in a straightforward example of scientific idealization (Potochnik 2017) we make our models more economical and explanatory by dismissing other cases as aberrations. Representational notions give us a good initial criterion for which cases to dismiss: ones in which the brain's causal structure does not mirror the environmental relationship of interest (at a fine or coarse grain) but, in the normative terms that representational thinking allows us to use, *mis*represents, *fails*, or gets its environmental target *wrong*. This normative terminology is common in idealization. E.g., we dismiss crystals that do not fit the prototypes described by our best mineralogy as *imperfect* (Polanyi 1966).

There is nothing stopping us from modeling misrepresentations if it is fruitful to. Pareidolia is an illuminating pattern of misrepresentation — a type of systematic failure that reveals interesting and relevant features of the causal structures we're modeling. Even though we see instances of pareidolia as misrepresentations, we care about capturing them in our models because we think they provide model-worthy information about the causal structures at issue (Liu et al. 2014). Likewise, some imperfect crystals may be worth our attention for various modeling purposes, even if most imperfect crystals aren't, for most purposes. So misrepresentations are not *necessarily* idealized away. Rather, thinking of something as a misrepresentation is a way to mark it as a deviation from the causal structure that is our main explanatory target. These deviations are dealt with on a case-by-case basis, but can often be idealized away at minor cost.

This account of misrepresentation gives questions about *what counts as a misrepresentation* a much smaller role than most other accounts, reflecting the minor role those questions play in cognitive science and especially neuroscience.[13] A 'fake' face, indistinguishable from a real one, would raise important questions on the standard account.[14] When we categorize it as a face, have we misrepresented it? If not, does that mean our representation is not of faces but of *face-like objects*? Does that mean that in cases of pareidolia, like when I see an electrical outlet as a face, I've correctly represented it? On my account, however, these questions fade away, leaving another: what do our categorizations of the 'fake' face tell us about the causal structures involved in face perception? If they mark some theoretically uninteresting deviation from the causal processes we're interested in, we can dismiss them as misrepresentations. If they involve causal processes we're interested in capturing, there's no need to dismiss them, and we may categorize these representations as correct or incorrect as it suits our modeling needs, i.e., as it suits our attempts to model the brain using structures defined over environmental entities, either including or not including the 'fake' face.

It will be apparent that much of the modeling process, and especially what counts as a misrepresentation and what we can idealize away, depends on our current understanding of the task domain and the brain's causal structure. And that understanding can change. If we begin to understand

[13] Of course, misrepresentations themselves, like illusions, play a major role in cognitive science. It's *questions about what counts as a misrepresentation* that don't. Though they do have *some* role in cognitive science. E.g., see the ecological psychologists who dispute the characterization of classical illusions as misrepresentations (Rogers 2022).

[14] Compare the discussion of fake *worms* — worm-shaped cardboard cut-outs — presented to a frog (Neander 2017).

face-discrimination as just a special case of expert discrimination (Kanwisher and Yovel 2006), we will model FFA and its role in face-discrimination differently, and we will dismiss a different set of cases as 'aberrations.' But this is a ubiquitous feature of science. The only problem would be if we had no grounds on which to support one understanding of the task over another. And we clearly do have those grounds, just like we do in any other case of scientific reasoning: we consider which understanding integrates well with our understanding of an organism's behavior more generally; which issues in models that integrate well with other models of the brain or models of other tasks; which requires less idealization or gets a better payoff for its idealizations; and so on. So, e.g., to justify modeling the hippocampus as representing an animal's *current* location, even though its activity is also correlated strongly with and can be modeled by an animal's *intended direction of movement* at an upcoming turn, it is enough to note that hippocampal activity correlates with intended direction only because the mice tend to *actually move* to one side when they are preparing to turn (Euston and McNaughton 2006). Then the general scientific criteria I've glossed will straightforwardly endorse an understanding of the hippocampus as representing (i.e., a strategy of modeling it with structures defined over) animals' actual locations, rather than their intended direction of movement.

I've given an account of what representational notions help neuroscience do. They help to project environmental structures onto the brain as models of how its causal structures bring about cognitive capacities. This strategy — representational explanation — has strict success conditions: the accuracy and success of the causal models. It is appropriate in some circumstances and not others, depending on both our explanatory goals and the features of the system it is applied to. And where it is appropriate, it has distinctive benefits: the tight and intuitive connections it provides between our models and the brain, and between the brain's fine-grained causal structures and the high-level capacities they support. Representational explanation also provides us with the resources to idealize the brain's causal structure in a way that is sensitive to our explanatory goals. And to understand how representational notions do all this, we haven't needed an account of what representation is. We've gotten by just describing *what representational notions help neuroscience* *do, and how*. That has been revealing regardless of the nature of any special property or relation those notions may refer to, and regardless of whether such a property or relation even exists.

4 Realism and methodological nominalism

My approach, as Steps 2* and 3* said, was to forget entirely about *what it is to be* a representation, or how that property should be defined, and instead elucidate representational explanation by discussing the way representational notions themselves figure into the explanatory economy of neuroscience. But it is commonly supposed that realism about representational explanation involves a commitment to “robust intentional properties” instantiated in parts of the brain (Ramsey 2021, p. 55), and to “the task of naturalising representation,” i.e., saying what exactly those intentional properties are (in naturalistic terms) (Sprevak 2013, p. 548). If that’s right, my approach could be dismissed as anti-realist, and I want to explain why that would be mistaken.

First, note that I haven’t argued that no property or relation NEURAL REPRESENTATION exists. I’ve argued that, in understanding how and why representational explanation works, we need not (and should not) concern ourselves with this property, or with the criteria for instantiating it. I’ll use the name “methodological nominalism” to capture the two essential aspects of this approach. The idea is to neglect representation itself — as a property, relation, or status that something can instantiate if it has the right qualities — just the way that a nominalist would neglect properties, relations, or universals corresponding to a predicate, instead focusing on other aspects of that predicate’s role in language and thought (e.g., Sellars 1960). And the approach is methodological in that the point is not to reject the property’s existence, but its relevance to a particular goal: understanding how and why neuroscientific explanations work.

Traditional scientific (anti-)realism is not a methodological view but a metaphysical one: a view about what there is, or about our ability to refer to it. Methodological nominalism is a view about where we should direct our attention if we want to understand a form of explanation. But even setting this aside, there are important differences between methodological nominalism and traditional scientific (anti-)realism. Traditional (anti-)realism is generally about one of two things: the existence of unobservable *entities*; or the truth of scientific *theories* (Chakravartty 2017). But, first of all, methodological nominalism has no qualms with unobservable entities in the brain, and it certainly has no problem with the (more or less) observable entities that we characterize in representational terms, like the brain's neurons and activities and structures and processes — and even its *representations*, as long as we mean the concrete activities and structures we're talking about when we use representational terms, and not a property or relation that those things instantiate and that philosophers puzzle over defining. Methodological nominalism is simply a commitment to understanding representational explanation without detouring through questions about *what representation is*. And while you could develop that view to be compatible with entity anti-realism, the version I've defended takes representational explanations to describe very real entities and processes in the brain, and this can only be understood as a form of entity realism.

Likewise, methodological nominalism has no qualms with the truth of neuroscientific theories. A methodological nominalist can accept (or deny) that representational explanations make true claims about the brain. We just don't think that one of these true claims is, "There is a set of things that have the status REPRESENTATION, and for something to have that status is for it to ...." And we don't think that, to understand the truths expressed in representational explanations, philosophers must make that kind of claim ourselves. By way of illustration, consider an ecologist explaining that a local change in fish populations resulted from a declining number of trees in the region. Only an especially unsophisticated theory realist would be defeated by the fact that *fish* and *tree* are such motley, jumbled, and indefinable categories that biologists tend to agree there's no such thing as a fish (e.g., Banister and Dawes 2005) and there's nothing it *is to be* a tree (e.g., Ridley-Ellis 2019). A sophisticated realist would see that all the explanatorily important facts can be retained without worrying about what it is to be a fish or tree. That is, the theory realist is committed to the reality of whatever the theory *itself* is committed to, and she has plenty of room to deny that theories using the notions of trees and fish are committed to the existence of the properties TREE and FISH. As long as she can find plausible scientific roles for those *notions* that don't implicate those *properties*, her realism is intact, and her understanding of ecology can be enriched by an account of the roles that the notions of trees and fish play.[15]

In the case of representational explanation, I've suggested that the explanatorily important facts (i.e., the ones that neuroscientific theories are committed to) have to do with the causal structure of the brain and the way that different components of that structure contribute to an organism's capacities. I've described a role for representational notions that can enrich our understanding of neuroscientific theories without implicating any property or relation, REPRESENTATION. And, while methodological nominalism could be developed as an anti-realist view, I have not questioned the truth of the explanatorily important facts. I have assumed that representational explanation provides (at least potential) explanations and understanding — not just prediction or control — of cognition by making (at least potentially) true claims about the brain's causal structure — not just by systematizing observations. This can only be understood as a form of theory realism, in stark opposition to traditional forms of anti-realism like instrumentalism.

But if the methodological nominalist can be a realist in both these senses, what do we make of the idea that realism involves a commitment to "robust intentional properties" instantiated in parts of the brain (Ramsey 2021, p. 55), calling for naturalization (Sprevak 2013, p. 548)? It is natural to think of nominalism as a kind of anti-realism about a property (though even in that case I've only defended a methodological version of the view). But this is not a troubling sort of anti-realism, nor a version of anti-realism that representation realists have refuted — or even challenged. It is not troubling because if I'm right about the way representational notions serve neuroscience, i.e., in a way that doesn't implicate a property or relation, REPRESENTATION, then it isn't a problem to neglect that property in our understanding of neuroscience. And it is unrefuted, in fact unchallenged, because the existing arguments for representation realism — even when they are

[15] This should also alleviate a potential worry: that methodological nominalism might apply too broadly. In many cases it will be difficult to argue that a theory's explanatorily important facts do not include the fact that a range of things fall into a special category. E.g., an area of science might aim to generate a taxonomy, or to systematically determine what kinds of things there are in a domain (perhaps physics or chemistry have these goals).

pitched as arguments *for properties* or *for the necessity of saying what representation is* — tend to be arguments for entity or theory realism.

E.g., Ramsey and Sprevak base their arguments on the idea that representations, as cognitive science understands them, have causal properties (Sprevak 2013, pp. 554-555; Ramsey 2021, p. 62). Likewise, Thomson and Piccinini (2018, p. 223) suggest that "the long-standing debate over representations should finally be settled" because representations have been observed and manipulated. These are arguments for entities. Finding that something has causal properties, or can be observed and manipulated, confirms the existence of *that something* (cf. Hacking 1983, p. 23), not the existence or relevance of a property corresponding to the notion under which you happen to be thinking of it (TREE, FISH, REPRESENTATION).[16] Other arguments support theory realism, e.g., by suggesting that it would be surprising if representational theories didn't accurately describe the structure of the brain Ramsey (2021, p. 60). No arguments for realism, to my knowledge, actually target the necessity of saying *what representation is* in order to understand representational explanation. So it is a mistake to think that realism about representational explanation, in the senses that have been defended and that appear to be most important, calls for a commitment to a special property or relation, REPRESENTATION, let alone an account of that property.[17]

[16] In case that point isn't obvious, imagine stubbing your toe on a tree. That should count as observation, manipulation, *and* causation. But it does not refute the point that there is nothing it *is to be* a tree, or the methodological version of that claim: that we won't learn much about ecology by asking what it is to be a tree (cf Brzović 2023).

[17] I've contrasted methodological nominalism with a broadly instrumentalist anti-realism because the latter has clear implications for our understanding of modeling practices like the ones I've been discussing. But a reviewer helpfully points out that I could also contrast methodological nominalism with *fictionalism*. Fictionalists about neural representation are rare, though Sprevak (2013) discusses the view, and there are fictionalists about mental states (Toon 2016). But a fictionalist would agree that neural activity needn't have the property REPRESENTATION for representational explanation to succeed. That, however, is where the similarities with methodological nominalism end. To make sense of representational explanation, the fictionalist would appeal to a fiction or pretense in which neural activity *does* have the property REPRESENTATION (cf. Walton 1993; Toon 2016). And the whole point of methodological nominalism was that this property is irrelevant. We can get by just fine asking what representational notions help us do, setting aside any property, real *or* fictional, they might refer to.

5 Upshots and objections

The view I've articulated gives a pragmatic answer to philosophical questions about how and why representational explanations work. They work by using representational notions to facilitate causal modeling, and they work (if and when they do) because representational notions facilitate modeling strategies that achieve particular explanatory goals. The view also gives pragmatic answers to the neuroscientist's questions. Which neural activity represents what? Which things should we understand as representations? Does FFA represent faces or some broader domain? The answer is that we should understand brain activities/structures in representational terms wherever it is helpful to apply the explanatory strategies that come with representational notions, and we should understand the brain as representing whatever domain provides the best models for it, given our interests in questions. As far as neuroscience is concerned, representation need not be a special property or relation that brain activity instantiates, or a privileged status it has. Thinking in representational terms is a *strategy* for modeling and understanding the brain.

All this followed from the decision to explore a different option at Step 2 of the three-step tactic. In this section I want to draw out two main advantages of that decision, and then consider some objections that were only partially addressed above. First, the advantages. An upshot of my view is that scientists using a workaday notion of representation can carry on, secure in the knowledge (as they presumably already are) that abstruse philosophical puzzles won't undermine their explanations. Philosophers and scientists studying *representational explanation*, on the other hand, can approach it as a form of explanation rather than a metaphysical commitment. And that has methodological implications. The standard approach has been limited to a priori analysis of the concept of representation and case studies of scientific explanation (e.g. Shea 2018; Cummins 1991; Ramsey 2007; Neander 2017). I don't mean to disparage that work. It has been illuminating, especially given the trend these past few decades towards scientifically well-informed case studies and detailed attempts to connect the property of representation to its scientific role. But the standard approach, and its focus on defining representation, does obscure the fact that we are fundamentally asking *how a certain form of explanation works*, and it obscures methods that could target that question more directly.

I'm thinking specifically about the psychology of explanation, perhaps best exemplified by Lombrozo and colleagues (Lombrozo and Carey 2006; Lombrozo 2009; Lombrozo et al. 2007; Lombrozo and Gwynne 2014). This work tries to understand different forms of explanation by asking where and why they tend to be applied, and, especially important for my purposes, what people *do*, cognitively, with the explanations, e.g., what predictions or generalizations they make given teleological as opposed to mechanistic explanations (Lombrozo 2009).[18] These methods are, of course, not applicable if we think that representational explanations work just by attributing a special property to a system or its parts. All that leaves room for is an investigation of the property, of what it might take to instantiate the property, and of what would be explained by something's instantiating it. But if we think of representational explanation along the lines I've described, as a non-metaphysically-committal contribution to scientific *thought*, then these other approaches become available to us. It is natural, and in principle straightforward, to apply the lessons, methods, and empirical paradigms used to study how concepts in general support explanation, to the question of how concepts in science do (Dubova and Goldstone 2023).

None of this is to say that we should do away with case studies (or a priori conceptual analysis, for that matter). My own argument has relied heavily on them. If you want to understand how a process like explanation works, it is useful to look carefully at examples of that process. In fact, case studies seem to fit more naturally into methodological nominalism's toolkit than into the standard approach's. Looking carefully at examples of scientific explanation should be informative about scientific explanation: about what it is, what it does, and how it works. My argument has banked on that. The standard approach is banking on something more complex: the idea that looking carefully at examples of scientific explanation will be informative about *the nature of a property or relation*

[18] The study of medical reasoning also illustrates this approach well (e.g., Goldszmidt et al. 2013, 2012). And note that these are not the typical methods of experimental philosophy of science. That field is generally committed to the standard approach, using experimental methods to uncover what scientists think representation (or some other property) *is*, or what qualities things must have to instantiate that property — not what role the notion of representation plays in scientific explanation (Favela and Machery 2023; Richmond 2023).

that the systems scientists study might instantiate, and that this in turn will be informative about our original questions concerning scientific explanation — what it is, what it does, and how it works. Methodological nominalism is simply a more direct approach to these questions, even when it's using the same case-study methodology. So if you want to give an account of *what it is to be* a neural representation, you need some reason to think that account will tell us more about how and why representational explanation works than an account that is, like the one I've given, explicitly about how and why representational explanation works. This is a challenge that proponents of the standard approach have not answered, and, to my knowledge, have not even been pressed to answer.

To summarize, the first advantage of methodological nominalism is the range of methods it offers philosophy of neuroscience, and the way it re-frames our current methods to answer our questions more directly. The next advantage has to do with the relationship between philosophy of neuroscience and neuroscience itself. Hacking (1983) suggests that where debates over realism (relatives of the current debate, though not identical to it) have been worthwhile, they have tended to occur in the context of pressing scientific debates. E.g., he suggests that anti-realism about Copernican theories was so important because of their conflict with Ptolemaic theories (Hacking 1983, p. 65). I take it this kind of connection to scientific concerns is at least a desideratum for philosophers of neuroscience. It needn't be, and philosophers who just want to play their own games with neuroscience's concepts are welcome to their pastime. But the philosophical debate over neural representation, and especially between representationalism and anti-representationalism, is generally taken to be relevant to neuroscience itself. So it is a problem that the standard approach has had limited impact on or connection with debates over representation *within* neuroscience. On reflection, it's clear why. Neuroscience's debates are generally not about what it is to be a representation, or whether some bit of neural activity meets the criteria.[19] They are debates over models and explanations: which are more accurate, predictive, simple, or revealing.

[19] Those debates do exist in neuroscience, as I discussed in Section 2, but they are exceedingly rare compared to the other sorts of debate I discuss here.

E.g., consider how Shenoy et al. (2013), along with the rest of the motor control community (e.g., Wang et al. 2022), understand the debate between representational and anti-representational (specifically: dynamical) approaches to motor cortex. This is not a debate about what it is to be a representation, or whether motor cortex meets the criteria. It is a debate about whether to model motor cortex as controlling motor activity through operations over neurons tuned to environmental and bodily variables, or as "generat[ing] motor commands by autonomous temporal evolution" through a state space (Wang et al. 2022, p. 796) (for a more thorough description of the debate, see Favela 2021). The anti-representationalism in that debate is a long way from the sort of anti-representationalism typical in philosophy, exemplified by Chomsky's arguments for eliminativism (Chomsky 1995) and Hutto & Myin's arguments against representationalism (Hutto and Myin 2014, section 6). Those arguments target the property of representation and some supposed incoherence or difficulty within it. On the view I've defended, the representational approach does not rely on any such property, and the debate between representationalism and anti-representationalism in philosophy should be precisely the same debate as the one in neuroscience: a debate over the right explanatory stance to take on some capacity or brain area, where the right explanatory stance is determined by whether representational notions, and the resources they introduce, generate accurate models that predict and generalize well, connect to their explananda, and so on. In fact, it's worth noting that on the account I gave in Section 3, representational explanation looks *just like* what representationalists and anti-representationalists are arguing over in motor cortex: the strategy of correlating brain activity with salient environmental variables to discover structures that model that activity. Philosophers like Hutto and Myin (2014) are right that there are plausible non-representationalist approaches. But their arguments against representationalism are misguided insofar as they challenge accounts, or the possibility of giving an account, of what representation is.

So in addition to providing an account of representational explanation and a range of methods that are well-suited to our goals as philosophers of neuroscience, methodological nominalism has another significant advantage: it puts philosophers in a position to contribute to genuine neuroscientific debates in a way that the standard approach, with its questions about how to define representation, does not. I now want to briefly address a pair of objections that were only partially addressed above, and then conclude by drawing a connection with another view of representation.

The objections point out contexts where methodological nominalism (or the way I've developed it) seems inappropriate. First, we might note the different sorts of debate the notion of representation figures into. The most striking are debates over *what represents what* — e.g., whether FFA represents faces or something else. A quick look at the FFA literature will show plenty of concern for this sort of question. Doesn't that imply that neuroscience is committed to there being a fact of the matter about what represents what, and criteria to settle that question? A longer look at the FFA literature, however, will show that these questions are part of a larger debate over how to model FFA, and especially which tasks our model must explain and which parts of the environment are therefore relevant (Kanwisher and Yovel 2006; Rhodes et al. 2004; Kasper et al. 2022; Schalk et al. 2017).[20] So the methodological nominalist can say the same thing here that I've been saying all along, even about debates involving very direct statements like, "FFA represents faces, and faces alone!" Debates over what FFA represents should be understood as debates over which capacities it supports and which environments are therefore relevant in modeling it, and how it supports those capacities and which models are therefore accurate. If FFA really is just involved in face perception, and if it performs face-perception with operations specific to faces, then a model of FFA will only have to explain that capacity and can probably get by relying on structures in the domain of faces. On my account, that means we should think of it as representing faces. But if FFA is involved in non-trivial ways in tasks aside from face-perception, and/or it performs face-perception with operations that are also involved in the perception of other domains, then the domain of faces will probably *not* provide sufficient models, and we will have to draw on other domains. On my account, that's exactly what we do when we think of FFA as representing more than faces.

The second objection points out that representational notions play a role in tasks aside from modeling, especially characterizing experimental targets, as Bechtel (2016) has emphasized. Doesn't

[20] Schalk et al. (2017) are especially clear about this, since they explicitly use hypotheses about what FFA represents interchangeably with hypotheses about what tasks FFA contributes to (p. 12286) and suggest that the point of representational hypotheses is to generate models of the causal structures that underlie cognitive capacities (p. 12289).

this show that my account is incomplete, focusing as it does on modeling and not experimentation? But my account actually *predicts* that representational notions will have a significant role in experimentation. If, as I have argued, representational notions characterize explanatorily important parts of a system's causal structure and connect them to explanatorily important environmental structures and variables, then how could representational notions *not* figure into experimentation — or, for that matter, prediction, the interpretation of data, and so on? These other tasks also target, or aim to manipulate, explanatorily important parts of a system's causal structure and their relationship to environmental structures and variables. To put it simply, experimentation, along with much of what scientists do, is informed by models and the understanding and expectations derived from them (not always sophisticated models, but Section 3 leaves room for representational notions in even simple verbal models).[21] In light of this, the discussion of experimentation in Bechtel (2016) does not contradict but complements the account I've given here.

I want to conclude by discussing the relationship between my view and another, called pragmatism or deflationism (Egan 2019, 2021; Mollo 2020; Cao 2022). Deflationism can be a bit hard to pin down. Sometimes it can seem that deflationists are taking the standard approach and addressing themselves to the question, *what is* neural representation?[22] Sometimes it is clearer that they are asking what role the concept of representation plays in cognitive science (e.g., Mollo 2020, p. 104). Regardless, they answer with a deflationary account: one that is distinctive in its sparseness and interest-relativity, and issues in a set of answers like those I gave in the first paragraph of this section. The view is often expressed, by deflationism's main champion, as the idea that ascriptions of neural representation "gloss" the *non-representational* characterizations of brain activity that constitute genuine scientific theories (Egan 2021). This gloss is supposed to serve pragmatic purposes like communication and, especially, linking mathematical theories of cognition to the capacities they

[21] Chirimuuta (2024) makes a similar point much more elegantly: "Like our hands, scientific models are both channels for knowing about things in the external world, and the means by which we manipulate those things."

[22] E.g., Cao (2022) argues that something's status as a representation depends on explanatory context (p. 18) and the way it is used by scientists (p. 15) — enough to count as a pragmatist or deflationist by anyone's lights. But she summarizes her view by saying that "two conditions must be met in order for some experimenter-detected activity correlated with a variable of interest *to count as a representation* of that variable: ..." (p. 16, emphasis mine).

explain (Egan 2010, p. 256). But it excludes the notion of representation from scientific "theory proper" (Egan 2021, p. 41).

That description should make it clear that a methodological nominalist will be sympathetic to deflationism. But I want to highlight a difficulty deflationists face. What that difficulty will bring out is that, first, methodological nominalism is pitched at a deeper level than deflationism, and second, deflationists would benefit significantly from adopting methodological nominalism.

So, consider the most stubborn objection to deflationism: that it collapses either to eliminativism (Neander 2015; Hutto and Myin 2021), or to a form of representation realism no different than the received view (Neander 2015; Ramsey 2021). Either the status of *being a representation* is "depreciated," "diminished," or trivialized by deflationists (Ramsey 2021, p. 74), in which case deflationism seems indistinguishable from eliminativism, which argues that the brain lacks this status except in a trivial sense (Hutto and Myin 2014). Or something's status as a representation is robust and non-trivial, in which case the received view seems perfectly capable of defining representation so as to accommodate the special features deflationists attribute to it (Ramsey 2021, pp. 74-76). Either way, the insights of deflationism — at least, the insights it might share with methodological nominalism — are lost.

But these objections[23] can't even get off the ground against methodological nominalism, because they object to an account of what it *is to be* a representation. Methodological nominalism cannot be mistaken for eliminativism because it doesn't "depreciate" the status of *being a representation* — it doesn't talk about that status at all. Methodological nominalists talk instead about the way representational notions serve neuroscience. And while it is open to us to conclude that, given the way they serve neuroscience, they should be eliminated, it is also open to us to endorse their use. And methodological nominalism cannot be mistaken for an exotic version of the received view for the same reason: it does not propose a special definition of representation — it rejects the attempt to give a definition at all.

[23] Along with many related ones. For quite a list, see Neander (2015) and Ramsey (2021).

The points of agreement between deflationism and methodological nominalism remain significant, especially given the way deflationists emphasize scientific goals and contexts (see Cao 2022, especially). But if the deflationist is still asking what representation is, or what it takes to count as a representation, they are subject to the objections I just described, and methodological nominalism provides the solution: we should ask how representational notions serve neuroscience, not what representation is. And if the deflationist already takes themself to be asking those questions, they need a principled way to ground them and explain their significance — to show exactly what makes those questions (and, a fortiori, their answers) not only worthwhile, but more significant and revealing than the standard approach's. And that has been the whole point of methodological nominalism: to show that, and why, questions about *what representation is* are irrelevant to our understanding of representational explanation, and to motivate a better set of questions.

To summarize, methodological nominalism is pitched at a deeper level than deflationism, and provides a philosophical foundation that deflationists (and others) can rest their views on. And as long as deflationists are contending with objections like the ones I've discussed, they would benefit significantly from building on that foundation. So I'm hoping deflationists will consider this an invitation. They have taken a significant step forward from traditional philosophical approaches to neuroscientific explanation. The invitation is to extend that step farther, to join the methodological nominalist in making a more fundamental point, and to embrace Step 2* as the explicit starting point for understanding representational explanation. From that starting point, philosophers of neuroscience (deflationist or otherwise) will have a methodologically sound basis for a philosophically illuminating account of representational explanation — and one that connects fruitfully to neuroscientific debates themselves.

References

Baker, B., Lansdell, B. and Kording, K. P. (2022), 'Three aspects of representation in neuroscience', *Trends in Cognitive Sciences* **26**(11), 942–958.
URL: *https://doi.org/10.1016/j.tics.2022.08.014*

Banister, K. E. and Dawes, J. (2005), Fish, What is a?, *in* A. Campbell and J. Dawes, eds, 'The Encyclopedia of Underwater Life'.

Bechtel, W. (2016), 'Investigating neural representations: the tale of place cells', *Synthese* **193**(5), 1287–1321.
URL: *http://dx.doi.org/10.1007/s11229-014-0480-8*

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L. and Kurth-Nelson, Z. (2018), 'What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior', *Neuron* **100**(2), 490–509.
URL: *https://doi.org/10.1016/j.neuron.2018.10.002*

Borghesani, V. and Piazza, M. (2017), 'The neuro-cognitive representations of symbols: the case of concrete words', *Neuropsychologia* **105**(June), 4–17.
URL: *http://dx.doi.org/10.1016/j.neuropsychologia.2017.06.026*

Brzović, Z. (2023), 'The Many Faces of Realism about Natural Kinds', *Journal for General Philosophy of Science* pp. 1–19.

Burnston, D. C. (2020), 'Contents, vehicles, and complex data analysis in neuroscience', *Synthese* **199**, 1617–1639.

Cao, R. (2022), 'Putting representations to use', *Synthese* **200**(151).

Carman, C. C., Thorndike, A. and Evans, J. (2012), 'On the pin-and-slot device of the antikythera mechanism, with a new application to the superior planets', *Journal for the History of Astronomy* **43**(1), 93–116.

Chakravartty, A. (2017), 'Scientific Realism'.

Chang, L. and Tsao, D. Y. (2017), 'The Code for Facial Identity in the Primate Brain', *Cell* **169**(6), 1013–1028.e14.
URL: *http://dx.doi.org/10.1016/j.cell.2017.05.011*

Chirimuuta, M. (2018), 'Explanation in computational neuroscience: Causal and non-causal', *British Journal for the Philosophy of Science* **69**, 849–880.

Chirimuuta, M. (2024), 'The Brain Abstracted – Overview and Precis'.

Chomsky, N. (1995), 'Language and Nature', *Mind* **104**(413), 1–61.

Craver, C. F. (2007), *Explaining the Brain*, Oxford University Press.

Cummins, R. (1975), 'Functional Analysis', *The Journal of Philosophy* **72**, 741–765.

Cummins, R. (1991), *Meaning and Mental Representation*, MIT Press.

Dennett, D. C. (1994), Cognitive Science as Reverse Engineering: Several Meanings of "Top Down" and "Bottom Up", *in* D. Prawitz, B. Skyrms and D. Westerstahl, eds, 'Logic, Methodology and Philosophy of Science IX', Elsevier Science, pp. 689–690.

Dubova, M. and Goldstone, R. L. (2023), 'Carving joints into nature: reengineering scientific concepts in light of concept-laden evidence', *Trends in Cognitive Sciences* **27**(7), 656–670.
URL: *https://doi.org/10.1016/j.tics.2023.04.006*

Edmunds, M. G. (2014), 'The Antikythera mechanism and the mechanical universe', *Contemporary Physics* **55**(4), 263–285.
URL: *http://dx.doi.org/10.1080/00107514.2014.927280*

Egan, F. (1999), 'In Defence of Narrow Mindedness', *Mind & Language* **14**(2), 177–194.

Egan, F. (2010), ‘Computational models: a modest role for content’, *Studies in History and Philosophy of Science* **41**, 253–259.

Egan, F. (2014), ‘How to think about mental content’, *Philosophical Studies* **170**(1), 115–135.

Egan, F. (2019), The nature and function of content in computational models, *in* M. Sprevak and M. Colombo, eds, ‘The Routledge Handbook of the Computational Mind’, Routledge, pp. 247–258.

Egan, F. (2021), A Deflationary Account of Mental Representation, *in* J. Smortchkova, K. Dolega and T. Schlicht, eds, ‘What are Mental Representations?’, Oxford University Press, New York.

Eliasmith, C. and Anderson, C. H. (2003), *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*, MIT Press.

Euston, D. R. and McNaughton, B. L. (2006), ‘Apparent encoding of sequential context in rat medial prefrontal cortex is accounted for by behavioral variability’, *Journal of Neuroscience* **26**(51), 13143–13155.

Favela, L. H. (2021), ‘The dynamical renaissance in neuroscience’, *Synthese* **199**(1-2), 2103–2127.
URL: *https://doi.org/10.1007/s11229-020-02874-y*

Favela, L. H. and Machery, E. (2023), ‘Investigating the concept of representation in the neural and psychological sciences’, *Frontiers in Psychology* **14**(1165622), 1–13.

Goldszmidt, M., Minda, J. P. and Bordage, G. (2013), ‘Developing a unified list of physicians’ reasoning tasks during clinical encounters’, *Academic Medicine* **88**(3), 390–397.

Goldszmidt, M., Minda, J. P., Devantier, S. L., Skye, A. L. and Woods, N. N. (2012), ‘Expanding the basic science debate: The role of physics knowledge in interpreting clinical findings’, *Advances in Health Sciences Education* **17**(4), 547–555.

Hacking, I. (1983), *Representing and Intervening*, Cambridge University Press.

Hutto, D. D. and Myin, E. (2014), ‘Neural representations not needed - no more pleas, please’, *Phenomenology and the Cognitive Sciences* **13**(2), 241–256.

Hutto, D. D. and Myin, E. (2021), Deflating Deflationism about Mental Representation, *in* J. Smortchkove, K. Dołęga and T. Schlicht, eds, ‘What are Mental Representations?’, Oxford University Press, chapter 3, pp. 79–100.

Kanwisher, N. and Yovel, G. (2006), ‘The fusiform face area: A cortical region specialized for the perception of faces’, *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**(1476), 2109–2128.

Kasper, V., Konkle, T. and Livingstone, M. (2022), ‘The neural code for ’face cells’ is not face specific’, *Arxiv* .
URL: *https://www.biorxiv.org/content/10.1101/2022.03.06.483186v1*

Kriegeskorte, N. and Diedrichsen, J. (2019), ‘Peeling the Onion of Brain Representations’, *Annual Review of Neuroscience* **42**, 407–432.

Liu, J., Li, J., Feng, L., Li, L., Tian, J. and Lee, K. (2014), ‘Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia’, *Cortex* **53**(1), 60–77.
URL: *http://dx.doi.org/10.1016/j.cortex.2014.01.013*

Lombrozo, T. (2009), ‘Explanation and categorization: How "why?" informs "what?"’, *Cognition* **110**(2), 248–253.
URL: *http://dx.doi.org/10.1016/j.cognition.2008.10.007*

Lombrozo, T. and Carey, S. (2006), ‘Functional explanation and the function of explanation’, *Cognition* **99**(2), 167–204.

Lombrozo, T. and Gwynne, N. Z. (2014), ‘Explanation and inference: Mechanistic and functional explanations guide property generalization’, *Frontiers in Human Neuroscience* **8**(700), 1–12.

Lombrozo, T., Kelemen, D. and Zaitchik, D. (2007), ‘Inferring Design: Evidence of a Preference for Teleological Explanations in Patients With Alzheimer’s Disease’, *Psychological Science* **18**(11), 999–1006.

Marchant, J. (2008), *Decoding the Heavens: Solving the Mystery of the World’s First Computer*, Random House, London.

Mekik, C. S. and Galang, C. M. (2022), ‘Cognitive Science in a Nutshell’, *Cognitive Science* **46**(8).

Mollo, D. C. (2020), ‘Content Pragmatism Defended’, *Topoi* **39**(1), 103–113.
URL: *http://dx.doi.org/10.1007/s11245-017-9504-6*

Moser, E. I., Moser, M. B. and McNaughton, B. L. (2017), ‘Spatial representation in the hippocampal formation: A history’, *Nature Neuroscience* **20**(11), 1448–1464.

Neander, K. (2015), Why I’m not a Content Pragmatist, *in* ‘The 2015 Minds Online Conference—the Brains Blog’.

Neander, K. (2017), *A Mark of the Mental*, MIT Press.

Palmer, S. E. (1978), Fundamental Aspects of Cognitive Representation, *in* E. Rosch and B. Lloyd, eds, ‘Cognition and Categorization’, pp. 259–303.

Polanyi, M. (1966), *The Tacit Dimension*, Doubleday, Garden City, N.Y.

Poldrack, R. A. (2020), ‘The physics of representation’, *Synthese* **199**, 1307–1325.

Potochnik, A. (2017), *Idealization and the Aims of Science*, University of Chicago Press, London.

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

Ramsey, W. (2021), Defending Representation Realism, *in* J. Smortchkove, K. Dołęga and T. Schlicht, eds, ‘What are Mental Representations?’, Oxford University Press, chapter 2, pp. 55–78.

Rhodes, G., Byatt, G., Michie, P. T. and Puce, A. (2004), ‘Is the Fusiform Face Area Specialized for Faces, Individuation, or Expert Individuation?’, *Journal of Cognitive Neuroscience* **16**(2), 189–203.

Richmond, A. (2023), ‘Commentary: Investigating the concept of representation in the neural and psychological sciences’, *Frontiers in Psychology* **14**.

Richmond, A. (n.d.*a*), ‘How computation explains’, *Mind & Language (forthcoming)* .

Richmond, A. (n.d.*b*), ‘What really lives in the swamp? Kinds and the illustration of scientific reasoning’, *Arxiv* .
URL: *https://philsci-archive.pitt.edu/id/eprint/22874*

Ridley-Ellis, D. (2019), ‘Wood you know a tree if you saw one’.
URL: *https://onlinevideo.napier.ac.uk/Play/15673!*

Rogers, B. (2022), 'When is an illusion not an illusion? An alternative view of the illusion concept', *Frontiers in Human Neuroscience* **16**(August), 1–13.

Schalk, G., Kapeller, C., Guger, C., Ogawa, H., Hiroshima, S., Lafer-Sousa, R., Saygin, Z. M., Kamada, K. and Kanwisher, N. (2017), 'Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain', *Proceedings of the National Academy of Sciences of the United States of America* **114**(46), 12285–12290.

Seiradakis, J. H. and Edmunds, M. G. (2018), 'Our current knowledge of the Antikythera Mechanism', *Nature Astronomy* **2**(1), 35–42.
URL: *http://dx.doi.org/10.1038/s41550-017-0347-2*

Sellars, W. (1960), 'Grammar and Existence: A Preface to Ontology', *Mind* **69**(276), 499–533.

Shagrir, O. (2001), 'Content, Computation and Externalism', *Mind* **110**(438), 369–400.

Shea, N. (2018), *Representation in Cognitive Science*, Oxford University Press.

Shenoy, K. V., Sahani, M. and Churchland, M. M. (2013), 'Cortical Control of Arm Movements: A Dynamical Systems Perspective', *Annual Review of Neuroscience* **36**(1), 337–359.

Shepard, R. N. (1984), 'Ecological Constraints on Internal Representation: Resonant Kinematics of Perceiving, Imagining, Thinking, and Dreaming', *Psychological Review* **91**(4), 417–447.

Sprevak, M. (2010), 'Computation, individuation, and the received view on representation', *Studies in History and Philosophy of Science* **41**, 260–270.

Sprevak, M. (2013), 'Fictionalism about Neural Representations', *The Monist* **96**(4), 539–560.

Thomson, E. and Piccinini, G. (2018), 'Neural Representations Observed', *Minds and Machines* **28**(1), 191–235.
URL: *https://doi.org/10.1007/s11023-018-9459-4*

Tolman, E. C. (1948), 'Cognitive Maps in Rats and Men', *The Psychological Review* **55**(4), 189–208.

Toon, A. (2016), 'Fictionalism and the Folk', *The Monist* **99**(3), 280–295.

Walton, K. L. (1993), 'Metaphor and Prop Oriented Make-Believe', *European Journal of Philosophy* **1**(1), 39–57.

Wang, T., Chen, Y. and Cui, H. (2022), 'From Parametric Representation to Dynamical System: Shifting Views of the Motor Cortex in Motor Control', *Neuroscience Bulletin* **38**(7), 796–808.
URL: *https://doi.org/10.1007/s12264-022-00832-x*

Waters, C. K. (2019), 'An Epistemology of Scientific Practice', *Philosophy of Science* **86**(4), 585–611.