What is a Theory of Neural Representation For?

[Redacted for blind review]

Abstract

This paper explores the way representational notions figure into cognitive science, with a focus on neuroscience. Philosophers have a way of skipping over that question and going straight to another: *what is neural representation?* The way representational notions figure into cognitive science is not forgotten — the phrase "neural representation" usually means "representation as cognitive science understands that notion." But eliding this phrase allows philosophers to focus more squarely on an account of neural representation itself. I argue that the wrong part of the question has been elided. Our ultimate questions, as philosophers of cognitive science, are about the function and epistemology of cognitive scientific explanations — in this case, explanations using representational notions. To answer those questions it is essential to understand the role the notion of representation plays in cognitive science — what it enables scientists to do or explain, and how — but not necessarily important to understand the nature of a property, NEURAL REPRESENTATION, that notion might pick out. I describe this approach, argue that it is a scientifically sensitive form of realism that philosophy of neuroscience can benefit from, and use it to give an account of representational explanation. Specifically, I propose that representational notions help us construct and understand models of the brain's causal structure, and that we can see how they do this by examining their role in scientific cognition, i.e., without debating the nature of any property they might refer to.

1 Introduction

Representational notions figure heavily in our understanding of the brain. Neuroscience in particular tells us that the brain supports navigation by representing spatial properties (Behrens et al. 2018), recognizes objects by representing their various features (Chang and Tsao 2017), supports language use by representing word meanings (Borghesani and Piazza 2017), and so on. So a central question in the philosophy of neuroscience has become, *what is neural representation?* What is this property[1] that the notion of representation, as it's used by cognitive scientists and especially neuroscientists, refers to? 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 the philosophy of neuroscience, it is not a *fundamental* one: it does not hold its central position because of the intrinsic interest of neural representation, but because of its context in a broader philosophical inquiry. We want to understand how and why neuroscientific explanations work; *that's* why we are driven to consider the properties they refer to — in this case, neural representation.[2] The main contribution of this paper will be to argue for a way of understanding neuroscientific explanation — representational explanation in particular — that does not detour through the debate over what neural representation *is*. Instead, I'll illuminate the role that representational *notions* play in the explanatory economy of cognitive science and especially neuroscience: how they help cognitive science achieve its explanatory goals. I'll especially stress the way representational notions help us construct and understand models of the brain's causal structure.

I'll start with some examples of representational explanation in section 2. I'll then illustrate the standard philosophical approach to representational explanation in section 3, noting the emphasis it puts on an account, definition, or metaphysics of the property NEURAL REPRESENTATION. In section 4 I'll outline my own approach, and in section 5 I'll use that approach to build an account of representational explanation, before discussing some objections in section 6 and concluding.

2 Examples: place cells and the fusiform face area

Many organisms have a remarkable capacity to navigate their environments, avoid obstacles, find remembered destinations, and travel home from new places along efficient paths. Our current

[1] I'll refer to the property of representation, though it can also be understood as a relation.

[2] This is why we ask *what is neural representation* and not (e.g.) *what is existential humor?*

understanding of how the brain supports spatial 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 "exhibit place-dependent activity independently of the animal's behavior or the task that it is performing" (Moser et al. 2017, p. 1448); that is, they respond selectively to locations in the environment. Together they tile the animal's environment, each representing its own preferred location (Moser et al. 2017, p. 1449). And they are well-suited to play a role in the kind of path integration algorithms that would support navigation, since they seem to combine information about the distances an animal has traveled in different directions (from collections of neurons that represent distance and 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 in its environment from representations of its previous movement directions, and distances.

Another capacity of many organisms is the ability to recognize and distinguish between faces (Kanwisher and Yovel 2006). In primates, 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, or as the particular faces they are, from a number of other representations: of face-parts (eyes, mouth, nose), of the spatial layout of those parts, and of 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* the 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, but the point is that explanations of face perception and spatial navigation are shot through with representational notions. What do these notions contribute? The explanation-sketches above show that representational notions *privilege* certain relationships — between place cells and an animal's current location, between the FFA and faces, etc. It's not just that some neural activity is correlated with faces or places, or carries information about them, or responds preferentially to them. The neural activity has those relationships with other things that we do not understand it as representing. E.g., place cell activity is correlated with an animal's movement *intentions* as well as its current location: place cells tend to fire before an animal changes direction, and their firing is correlated with the direction it ends up moving (Euston and McNaughton 2006). FFA activity is also famously correlated with many things aside from faces (Rhodes et al. 2004), to the point that there is a case to be made that the FFA is actually best understood as representing non-facial features (Kasper et al. 2022). But what's important is that no one claims the relevant structures or activities represent *just whatever* they're most correlated with, or carry the most information about, or so on[3] — though these are common and useful targets for experimentation (Baker et al. 2022). When we talk about representation, we're not talking about a straightforward physical, formal, or statistical relationship between the brain and part of the environment. We are, again, privileging some such relationships over others.

3 Questions about representation

Neuroscientists ask all kinds of questions about representations. Which parts of the brain represent? What do they represent? And what neural structures implement the representations? Philosophers tackle more fundamental questions about representational explanation *qua* mode of explanation. What is the function and epistemic status of representational explanations? How do they work? And *why* do they work — why are they successful, if and when they are?

[3] Famously, that kind of claim runs straight into the disjunction problem (see, e.g., Fodor 1987).

The *standard approach* in philosophy holds that both sets of questions are best approached via an account of the nature or metaphysics of representation, through some kind of definition. This is perhaps best understood as a simple, and quite sensible, three-step tactic:

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, distinctiveness, and perhaps success, it is important to understand how and why these explanations work.

Step 2 Provide a plausible skeleton answer to those questions: *the explanations work by attributing representations to the brain.*

Step 3 Put flesh on this answer by saying what precisely the explanations attribute to the brain when they attribute representations to it. That is, say what it is for something to be a representation.

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: how to define the property REPRESENTATION. And most philosophical work on the subject does precisely that: it aims to say which relationships between brain and environment[4] are representational, and why: what *makes* them representational. This is not just an answer to the philosopher's questions but the neuroscientist's too: a neural structure or a bit of neural activity will either satisfy the definition of *representation* or not; it will meet the criteria to be a representation, and a representation of x for any x, or not. That will tell us which parts of the brain represent, what they represent, which structures implement the representations, and so on.

This approach has generated illuminating work. Cummins' foundational book on mental representation is one example. He begins with the assumption that neuroscientists using representational

[4] Or between the brain and other activity in the brain — e.g., when we say some population of neurons represents the uncertainty in another population's representation. But nothing is lost for my purposes if we focus on brain–environment relations.

notions must use them to refer to some representation *relation*, which it is up to philosophers to define, in order to set foundations for neuroscience:

> 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 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.[5] (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 (the first sentence), to Step 3, the question of what precisely the property or relation of representation is (the second and third sentence). Given a different assumption at Step 2, this transition would be a clear non-sequitor. It is only because we assume the explanations work by attributing a property, REPRESENTATION, to the brain that we think we need an account of what precisely that property (or relation) is. This assumption is dominant in recent work as well. E.g., Shea begins his account of neural representation by moving, just like Cummins, straight from the existence of representational explanations to puzzles about what exactly the property of 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. This is an undoubted lacuna in our understanding, a void hidden away in the foundations of the cognitive sciences. (Shea 2018, p. 5)[6]

[5] I take it Cummins is addressing the question I posed in the introduction, about *neural* representation in particular, because so much of his discussion is about the fourth item on his list of "things that can be mental representations" (Cummins 1991, 2) — namely, "(actual) neurophysiological states" (Cummins 1991, 6).

[6] Both Cummins and Shea are after a 'naturalistic' metaphysics of representation, or a definition that reduces the notion of representation to something more fundamental, but that isn't the part of their view I'm targeting. Someone

As I've suggested, I plan to doubt that lacuna. I will provide a different answer at Step 2, and that will call for a different approach to Step 3 as well. But I want to note two things before moving on.

First, even philosophers who emphatically agree that the important questions in this area are about how and why representational explanations work still tend to take the standard approach. Ramsey (2007), e.g., frames his view of representation in terms of the role that cognitive science needs representations to play, or the "job description" that neuroscientists set for representations, which a bit of neural activity has to satisfy to count as a representation (Ramsey 2007, 24-25). But as that description makes clear, Ramsey still thinks that to understand representational explanation we need to investigate the property, standard, or description that something must instantiate, meet, or satisfy in order to *be a representation*. As he says, pointing to the same lacuna as Shea, his goal is to show "what it means for something to function as a representation in a cognitive system" (Ramsey 2007, 188). Even more explicitly, his goal is to analyze "the sort of physical conditions and relations that have been assumed to bestow upon an internal state the status of representation," and to propose his own set of conditions (Ramsey 2007, 189). 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, and it is what my approach will abandon.

Second, 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 quick look at almost any neuroscience journal will show plenty of concern for representation*s*, but no concern for the kind of debates or objections that an account of the *property* of representation would have to tackle, like the question whether one's definition of representation includes things that are (arguably) not representations. But occasionally a neuroscientist will enter into the metaphysical debate, or at least frame their questions in the metaphysical terms philosophers have set. E.g., Eliasmith and Anderson set out to understand the nature and significance of *representational claims*, like claims that some area

who, e.g., thought that representation can't be defined in more basic scientific terms, but must be understood as the intrinsic possession of truth-conditions, would still be pursuing the standard approach as I've described it here, through the three-step tactic.

of the brain represents some property (Eliasmith and Anderson 2003, p. 5). That looks a lot like Step 1, as I described it above (leaving aside any difference between representational 'claims' and representational 'explanations'). They then follow through on the three-step tactic, assuming that to understand representational claims they must "determine the exact nature of the representation relation; that is, . . . specify the relation between, and representationally relevant properties of, things 'inside the head' and things 'outside the head"' (Eliasmith and Anderson 2003, p. 5). So my targets are not just philosophers. My targets are the philosophers and the rare neuroscientists who think that a definition or metaphysics of representation is a prerequisite for understanding representational explanation.

My own approach, instead, will revise the three-step tactic like so:

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, distinctiveness, and perhaps success, it is important to understand how and why these explanations work.

Step 2* Provide 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* Put flesh on this answer by saying what resources representational notions introduce, and how those resources serve neuroscience's explanatory goals.

If you really wanted to miss the point, you might note that one thing a notion can do to serve neuroscience's explanatory goals is to refer to a property: REPRESENTATION. But the point is that notions can do other things too, and that investigating those other things is a promising way to understand how and why representational explanation works.

I don't think it's necessary to 'make room' for a view like this by defeating all others (cf Chemero 2011, 3-16). So I won't. The account will either work, explain, illuminate, or it won't. To motivate the account it is enough to note that Step 2 is optional, and that Step 2* is another option. The particular benefits of taking Step 2* will be clearest when we've seen the account of representational explanation it results in. Before turning to representational explanation, though, let me say a bit more about what my approach is, and especially what it is not.

4 Methodological nominalism

My approach, as Steps 2* and 3* say, will be to forget entirely about the property of representation, and instead elucidate representational explanation through a discussion of the way *representational notions themselves* figure into the explanatory economy of cognitive science. Though this could be cashed out in different ways, I further assume that representational explanation works by using the notion of representation to introduce *conceptual resources* that serve neuroscience's explanatory goals. That means Step 3* is to describe those resources, and show how they serve those goals.

This means I am not arguing that a property of neural representation doesn't exist. I am not even arguing that the notion of *representation* fails to refer to such a property. I am arguing that *it is not by referring to such a property that representational notions serve neuroscientific explanation*. That is, in understanding representational explanation in neuroscience — in figuring out how and why it works — we need not (and perhaps should not) concern ourselves with such a property. For our purposes as philosophers of cognitive science, the property of neural representation can be ignored. The (somewhat clunky) name "methodological nominalism" is supposed to capture the two essential aspects of this approach: the idea is to neglect the *property* of representation, the way that a nominalist would neglect properties or universals corresponding to a predicate, and understand that predicate and our practices around it by appeal to different resources (e.g., Sellars 1960); and the approach is methodological in that the point is not to question the property's existence but its relevance to a particular goal — understanding how and why neuroscientific explanations work.

Methodological nominalism is not a traditional scientific anti-realism. Traditional anti-realism is a metaphysical view, a view about what there is, or perhaps about our ability to refer to it. Methodological nominalism is not. Even if traditional realism was committed to the existence of a property of representation (on which more in a moment), this would not bring it into conflict with methodological nominalism: the methodological nominalist argues only that this property, whether it exists or not, has no role in answering our questions about representational explanation. If the methodological nominalist is successful, she will have answered those questions without detouring through debates about the existence and nature of the property of representation. So, although she will have cast doubt on any approach that makes the resolution of those debates a central task, she will remain neutral on the existence of the property of representation.

Even setting aside this difference, there are important differences between methodological nominalism and traditional scientific anti-realism. Anti-realism is generally one of two things: a view about the existence of unobservable *entities*; or a view about the truth of scientific *theories* (Chakravartty 2017). Even if we ignored the *methodological* part of methodological nominalism, it would be about neither: it has no qualms with the *stuff* of the brain (even when that stuff is the kind of non-observational stuff with which entity anti-realism is concerned), just with what we are committed to when we characterize that stuff. The neurons and activities and structures and processes in the brain are all relevant to cognitive science — *as are the representations*, so long as we mean the concrete stuff and causal structures we're talking about when we use representational notions, and not a property, REPRESENTATION, that this stuff instantiates and that philosophers puzzle over. Methodological nominalism is consistent, and fits well, with a paradigmatic entity realism like Hacking's (1983, p. 23) as opposed to a traditional anti-realism like van Fraassen's (1980).[7] When we say "if you can spray positrons, they're real," we're committing to the entities, concreta, stuff, that we call "positrons." That leaves it open to either accept or deny that a property, POSITRON-HOOD, exists. And, more to the point, it leaves it open whether our philosophical understanding of physics depends on defining that property or giving an account of its nature. A methodological nominalist about positrons (I'm not endorsing the view!) wouldn't question

[7] So methodological nominalism is not opposed to the entity realism that some philosophers, like Thomson and Piccinini (2018) and Bechtel (2016), target. It is essential to distinguish between entity anti-realism and nominalism to avoid confusion. E.g., see the arguments for realism from the fact that representations have causal properties in Ramsey (2021, 62) and Sprevak (2013, 554-555). These are sound arguments for an uncontroversial sort of entity realism, but not for anything more.

the legitimacy of claims like "the positron left such-and-such a trace" — she would question the necessity, for understanding this claim, of investigating the property POSITRON-HOOD.[8]

The above also means that methodological nominalism poses no challenge to what cognitive science says about the brain and its causal structure. In other words, methodological nominalism has no quarrel with *theory* realism. Methodological nominalists can be representationalists.[9] We can be happy for representational theories to be true, representational explanations to be explanatory, and representational models to be accurate, because we think their truth/etc. need not be understood in terms of a property of representation — we don't think it's part of their content that anything instantiates such a property. What's under scrutiny is not the *truth* of representationalism, but its *commitments*.

To make one more distinction, methodological nominalism is not a Dennettian sort of instrumentalism. For one thing, Dennett's view explicitly tries to say what something must do or conform to in virtue of which it counts as, or can be called, a representational system (Dennett 1988, 496 and passim). As I've described in this section and the last, this is emphatically opposed to methodological nominalism.[10] A more important difference is that Dennett's instrumentalism is quite cavalier about the structure of the brain: even where representational explanations are successful, they tell us "nothing ... about the ultimately mechanical details" of our brains (Dennett 1988). This is not the methodological nominalist's approach: we can be realists about exactly those mechanical details, and, in fact, the view I'll describe in the next section has it that *the main function* of representational notions is to help us describe those details.[11]

[8] More on how far methodological nominalism might extend in section 6.

[9] We can be anti-representationalists too. Methodological nominalism is a view about how a particular type of explanation works, and why it works *if and when* it does. We can describe representational explanation and go on to either endorse or reject it. For ease of exposition I'll assume a broadly representationalist approach, but I'll discuss anti-representationalism in section 6.

[10] It's not clear that this is an essential feature of Dennett's view, though it undeniably is *a* feature (again, see Dennett 1988, 496 and passim).

[11] On my understanding, Dennett wants to detach representation from mechanical detail because he's concerned with propositional attitude-ascriptions in psychology and folk psychology, and the possibility of inferring, from their success, a Language of Thought instantiated in the brain (Dennett 1988, 497). So our difference is largely due to our explananda: Dennett just isn't concerned with representational notions *in neuroscience*, where they are primarily used to describe mechanical details and causal structures in the brain, as I'll discuss in the next section.

So we can remain committed, as I do in the following and as traditional anti-realists and instrumentalists do not, to the claim that models in cognitive science, including ones couched in representational terms, furnish explanations and understanding (not just prediction or control) of cognition by capturing the brain's internal causal structure (not just by systematizing or otherwise describing observations). We can be straightforward realists about cognitive science even as it uses the notion of representation, and as it says things like "the brain derives representations of faces from representations of facial features."[12] I'll have more to say about this in the final section — for now, on to the positive account of representational explanation.

5 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. 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 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 contribute to your engineering project? It helps you to literally impose the structure of the logical functions (defined over mathematical objects) onto the causal structure of the gates by connecting the gates so that their causal structure mirrors that logical structure. Another way of putting this is that the logical structure acts as a model, and in thinking of the gates as representing parts of that model, you are *cognitively connecting them* to the model to help you impose, on their causal structure, the formal structure of the model. As you build the computer you will think about its inputs and outputs as representing elements of the domain the model is defined over (1s and 0s), and you will talk about the system in terms of the model and its domain. 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

[12] To address one more form of anti-realism, methodological nominalists are not fictionalists (Sprevak 2013). We are not saying that representational explanations talk about a fiction where the property of representation exists; we're saying that representational explanations don't talk about that property at all.

the system not in terms of its own electrical and physical properties, but literally *in terms of* the model you think of it as representing, and whose structure you want it to mirror.[13]

But what if you weren't 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? I submit that you would 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 mathematical functions, etc., outputting numbers, truth–values, and so on, in response the same given as inputs — again, *descriptions literally in terms of a mathematical or logical model* at a coarse grain. And you would describe the internal causal structure of the computer with algorithms that compute those functions, describing structures as AND-gates or electrical impulses as 1s and 0s, e.g. In other words, you would talk about the internal processes as well as the inputs and outputs as representing different components of the model, and this would provide that same link between physical system and model that we saw in the forward-engineering example.

To summarize these examples, understanding the computer's inputs and outputs as representing mathematical or logical entities means understanding them in mathematical or logical terms (literally: *in that terminology*; using the concepts or notions those terms express), and understanding the computer in terms of a function over those entities. And this provides an intuitive way of using the relevant mathematical or logical terms, and the formalisms they figure into, to describe its internal causal structure: in terms of algorithms that would compute the mathematical function. This not only identifies a space of potential models, but provides an intuitive link between the causally

[13] 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 *actually 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. It is easy to imagine the problems of indeterminacy that would result if you took the fact that we can use x to model y to constitute a representation relation between x and y (Sprevak 2010; Shagrir 2001).

relevant parts of our target system and the aspects of the model they should correspond to — i.e., the parts we think of them as representing — if that model is to be accurate and explanatory.

It is important that the models need not be defined over abstract or mathematical objects. Compare an actual computer found on a beach (or close enough): the Antikythera mechanism, commonly known as the first computer. This ancient Greek device calculated astronomical relationships. Since it was discovered, its inputs and outputs, as well as its internal causal structure, have been understood representationally and thereby modeled using structures defined over astronomical entities (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 domain the models are defined over (Carman et al. 2012).[14] Understanding the mechanism as representing astronomical entities and relations allows us to talk about it in terms borrowed from that domain, and we talk about it in those terms in order to project structures from that domain onto the mechanism as models of its causal structure. Relationships between astronomical entities model relationships between input and output in the mechanism, and between individual components within the mechanism.

Here again, representational thinking helps us create models. It helps us connect them to our target systems by thinking of those systems in terms of the models. And it ensures that our models (if they are accurate) clearly explain the system's capacities: there's no chance we lose track of how a model explains a system's capacities (to add, or to track astronomical relationships) because the model is specified precisely in terms of those capacities (the numbers added, the astronomical relationships tracked).

[14] This is even more pronounced in popular treatments, where, e.g., a function 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).

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 try to understand a complex system*. They allow us to impose structures from the 'represented' domain onto the 'representing' system as models, not just in engineering a system but in reverse-engineering 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 the goal of cognitive science: to reverse-engineer the brain by constructing models of its causal structure insofar as that structure supports cognitive capacities (Dennett 1994).[15] Those capacities, like in the cases above, are generally understood as abilities to produce certain environmentally-defined outputs as responses to environmentally-defined inputs, stimuli, or states of affairs more broadly.[16]

The FFA, e.g., is understood as taking low-level environmental features as input, and giving categorizations of entities as faces or particular faces as output. Just as in the mathematical cases, or the case of the Antikythera mechanism, there is a relationship between the environmentally-described inputs and outputs — not a relationship between addends and their sum or between the motion of the sun and the motion of the moon, but a relationship between the low-level features of an object and its being 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 — it must mirror that same function in its causal structure; it must have a causal structure that is accurately modelled by the function from low-level features to something's status as belonging to the category *face* or *non-face*.[17] Thinking about the FFA as representing faces allows us to project that function, as well as algorithms that would compute it, onto the brain's causal structure. Just as we did with the computer, we are using representational notions to connect

[15] Compare Mekik and Galang (2022) for a related but more detailed description of this approach.

[16] Though this must include *internal* outputs (like new memories or subjective experiences) and inputs (like goals or stored memories) too.

[17] I'll discuss the ways a system can deviate from that structure shortly.

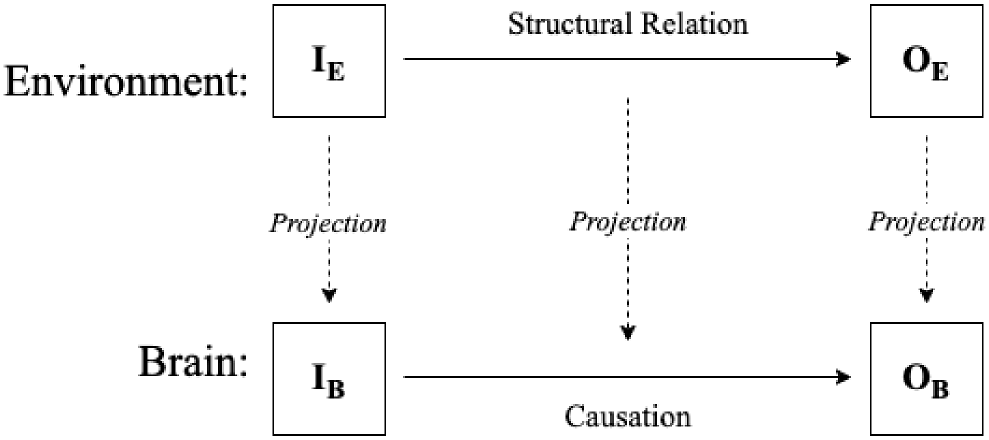

Figure 1: A schematic showing the relation between an input from the environment (I_E) and output to the environment (O_E), and the projection of their relationship, and the algorithms that would transition from one to the other, onto the brain as a model of its causal structure insofar as it transitions from an internal state (I_B) corresponding to the input to an internal state (O_B) corresponding to the output.

the brain to its task domain and to project structures from that domain onto the brain as models of its causal structure (see Figure 1).[18] Then we can test for those structures in the brain, just as we would test for causal structures in a computer after we had modeled them in logical or mathematical terms.

The FFA may be a simpler case than most (at least on my treatment of it here) but the same story can be told elsewhere. In the case of navigation, we model the brain with a function from sensory stimulation or previous states of the environment to an action or a future state of the environment. E.g., consider a mouse that reliably finds the most efficient path home from a foraging trip. There 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 discussed in Section 2, the path home tends to be about the most efficient one available. Because it moves from the former to the latter environmental entities or states — from a set of distances and directions travelled to a new route taken — the animal must move between internal states corresponding to those environmental ones. It must have a causal structure that can be modeled with reasonable accuracy by a function from the distances and directions travelled to the most efficient path home, and by some algorithm or process that computes that function.

[18] Compare Egan (2014) and Cummins (1991) in relation to the figure, and see section 6 for more on Egan's view.

Let me make a caveat and a qualification. The caveat is that I intend the notions of computation and algorithm extremely broadly, as any description of a sequence of states for implementing some function.[19] The formal description could be a computer program, a dynamical equation describing evolution through a state space over time, a graph-theoretic description of node activity, a verbal description of transitions between states, etc.

The qualification is that representational explanation will not always be a good modeling strategy. In some cases, e.g., a pure dynamical model will be more appropriate, where the processes a system implements are modeled by dynamical equations that have little to do with the structure of the environment. 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. But a better model describes the overall dynamics of the Watt governor without detouring through its environment, except perhaps to describe its input and output in terms of vehicle speed or combustion rate. Where those inputs and outputs — defining the 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 the *internal structures* with algorithms whose stages or transitions are *themselves* defined over environmental entities.[20] E.g., we do not just model the brain as moving from states corresponding to low-level visual features to states corresponding to the perceived object's being 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're describing the internal processes in terms of fine-grained 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-level features. And we're using these

[19] Thanks to [redacted for blind review] for discussion on this point, though I think she will find this caveat much too brief.

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

relationships between features to model not just the brain's input–output functions, but the *steps* between input and output at a relatively fine grain.

With all this in mind, when is representational explanation likely to be a good modeling strategy? First, it is essential that our interest in the target system is to explain how it brings about capacities described as input–output pairings defined over some domain outside the system[21] (whether a mathematical or environmental domain) — e.g., the capacity to recognize faces, or to get from one place to another. Otherwise, external structures are unlikely to provide relevant models of the causal structures we're interested in. Representational explanation will be most useful when the target system is also complex, necessitating some strategy for navigating a large and complex space of possible models, and a strategy for clarifying and highlighting the models' explanatory connection to the target system's capacities. 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 being selected. It is because we faced selection pressure to navigate accurately that the hippocampus internalized causal structures recapitulating environmental structures, and this is at least part of the reason that it can be accurately modeled by those structures.

This kind of result is not, of course, what design or evolution always create. The Watt governor was designed, and could no doubt have been selected for. And neither evolution nor selection appear necessary for representational explanation to apply accurately and fruitfully (see Richmond n.d.). But as long as we understand representational explanation as a tool or modeling strategy, all the previous paragraph claims is that when you're dealing with a complex, evolved system, and your interest is to explain how it brings about capacities described in environmental terms, representational explanation is a tool you'll likely find useful.[22]

[21] Or outside the particular system component we're interested in.

[22] To put a finer point on it, this discussion of evolution is no capitulation to teleosemantics: evolution and design do not figure into definitions of representation; they do not even tell us how and why representational explanation works.

The conclusion of all this is that representational notions give us a way of identifying, and projecting onto the brain as models, environmentally-defined structures that might serve as good models of the brain's causal structure: structures that, if they did accurately model the brain, would explain its 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 purely electrical and physical terms.

As I've advertised, this is an account of what representational notions allow us to do, *not* an account of the property of representation or the representation relation. To drive this point home, note that my account would be utterly hopeless as the latter. Any system can be modeled by a huge variety of structures, especially if we allow ourselves some liberty carving up the system into parts (see Richmond 2022) or stating the constraints on the model. If the hippocampus represents every domain containing structures that could model it — even just domains with structures that could model it *extremely well* — 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 describing the property of representation, I have been describing the way representational *notions* help cognitive scientists in their explanatory and modeling tasks. I have proposed that representational notions do this not by referring to some property that their target systems must instantiate, but by providing tools for constructing and understanding models. The point is to answer the question of how and why representational explanation works without *even having to enter* 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 *are really* representations Ramsey (2007, 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. If the answer is yes, we should think about and model those systems using

They only elucidate some conditions under which it is likely to be fruitful, due to the *non-teleological* specification of how it works that I've given.

representational notions, and we will have a case of representational explanation. If the answer is no, we shouldn't and we won't.

It is also worth noting that because of all this, my account is only distantly related to isomorphism theories of representation (Cummins 1991; Ramsey 2007). Aside from the important difference I just noted, there is another: (approximate) isomorphism plays no role on my account, except insofar as what makes *any* explanation involving a model appropriate is, partly, an isomorphism between the model and its target system. What is distinctive about representational explanation is not the isomorphism that every other model-involving explanation involves as well. What is distinctive about representational explanation is the kind of conceptual resources it uses to construct and understand models. And what makes a representational explanation *appropriate* is not primarily an isomorphism between two systems — except, again, in the sense that this is required of every model-involving explanation. What makes a particular representational explanation appropriate is that it provides a useful and accurate model of causal structure. And what makes representational explanation in general an appropriate strategy is the set of considerations I described above, to do with the broader explanatory context and goals.

To round off this section, I want to turn to a potential worry — one that will also let me illuminate a feature of this account. Take the FFA again. The function from the low-level visual features of an object 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 that 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 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 and classify those failures. Face-recognition has some illuminating patterns of error (consider pareidolia or prosopagnosia) that we 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, 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 they would not offer explanatory gains sufficient to justify their complexity and the extra work involved in creating and using them. Nor would they 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. This is a straightforward case of scientific idealization (Potochnik 2017): to make our models more economical and explanatory, we dismiss certain aspects or instances of our target phenomenon as aberrations. Representational notions give us a good criterion for which cases to dismiss: ones that, on our understanding of the capacities we're studying, must be classified as misrepresentations, i.e., ones in which the brain's causal structure does not mirror the environmental function of interest (either at the input–output level or the finer-grained levels) but, in the normative terms that representational thinking allows us to use, *mis*-represents, *fails*, gets its environmental target *wrong*, or otherwise acts as it should not according to our model of it. 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).

Let me make three small points before ending this section. First, as I've indicated, there is nothing stopping us from including misrepresentations in our model if it is fruitful to include them. Pareidolia is an example of 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 of our purposes. So misrepresentations, on the account I've given, are not necessarily idealized away. But thinking of something as a misrepresentation is still a way to mark it as a deviation from the causal structure that is our main explanatory target; these deviations are then dealt with on a case-by-case basis, but can often be idealized away at minor cost.

Second, it is worth noting that misrepresentation, and veridicality conditions as a whole, end up with a much more minor role on this account than most others, reflecting the minor role they actually play in cognitive science and especially neuroscience. A 'fake' face, indistinguishable from a real one, would raise important questions on the standard account.[23] 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*? And what does that mean for pareidolia? Or can we leave these questions open, allowing for representational indeterminacy? If so, under what conditions is representation indeterminate? 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 structures, either including or not including the 'fake' face.

Third, it will be apparent that much of the modeling process, including what counts as a misrepresentation and what we can idealize away, depends on our current understanding of the task domain and of the brain's causal structure. And that understanding can change. If we begin to understand face-discrimination as just a special case of expert discrimination (Kanwisher and Yovel 2006), we will model the FFA and its role in face-discrimination differently, and the patterns of 'success' and 'failure' we identify will change as well. But it hardly needs stating that the proliferation of models isn't a problem; it's a ubiquitous feature of science. The problem would be if we had no grounds on which to support one understanding of the task over another. And we

[23] The more common discussion is of fake *worms* — worm-shaped cardboard cut-outs — presented to a frog (Neander 2017).

clearly do have that from sources common in scientific reasoning: we consider which understanding integrates well with our understanding of an organism's behavior more generally; which one issues in models that integrate well with other models of the brain or models of other tasks; which one requires less idealization or gets a better payoff for its idealizations; which one issues in models at the desired level of grain; 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 or the other of their corridor in preparation for the turn (Euston and McNaughton 2006). Then general scientific criteria will issue in a straightforward endorsement of modeling the hippocampus as representing (i.e., with structures defined over) an animal's current location, rather than its intended direction of movement.

Aside from the specific details of this account, what's important to take away is that nothing here requires a definition of the property of representation, or even the assumption that such a property exists. I've talked only about what representational notions allow us to do. They allow us to project environmental structures onto the brain to generate and understand models that are tightly and intuitively connected to our explananda, and to make principled idealizations of the brain's causal structure. The way they do this does not depend on the brain's structures or activities instantiating some property, REPRESENTATION. Looking at what representational notions help us do is revealing regardless of the nature of the property they may refer to, and regardless of whether it even exists. We may be able to learn all we want to know about representational explanation without ever discussing that property.

6 Upshots and objections

The view I've articulated gives a pragmatic answer to our 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 cognitive science's explanatory goals. The view also gives pragmatic answers to the neuroscientist's questions. Which neural activity represents? What does it represent? More generally, which things and relations are privileged as representational, and why? In themselves: none are. *We privilege* certain relationships because doing so helps us build models of the brain's causal structure. As far as neuroscience is concerned, representation need not be something the brain does, or a privileged relationship neural activity has to certain things. Representation is a notion that helps neuroscience model and understand 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 benefits of that decision, and defend it against some objections. First, the advantages. An upshot of my view, mentioned in the first section, 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. What about philosophers and scientists studying representational *explanation*? We can approach it *as a form of explanation*, rather than a metaphysical commitment. And thinking of representational explanation along these lines has an important methodological implication. 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). I don't mean to disparage that work. It has been illuminating, especially given the trend these past few decades towards detailed, careful, and scientifically well-informed case studies (e.g., see Neander 2017; Shea 2018). Any philosopher of cognitive science could learn a great deal from this work. But the standard approach, and its focus on the metaphysics of 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, as exemplified by Lombrozo and colleagues (Lombrozo and Carey 2006; Lombrozo 2009; Lombrozo et al. 2007; Lombrozo and

Gwynne 2014).[24] 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 are able to do, cognitively, with certain types of explanations, e.g., what predictions or generalizations they can make given teleological as opposed to mechanistic explanations (Lombrozo 2009). These methods are, of course, not applicable if we think that representational explanations work just by attributing the property of representation to a system or its parts. All that leaves room for is an investigation of the property we're attributing to the system. But if we think of representational explanation along the lines I've described, as a non-metaphysically-committal contribution to scientific practice, and especially scientific *cognition*, 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 explanations *in general* support cognition to the question of how explanations *in science* do.

None of this is to say that we should do away with case studies (or a priori conceptual analysis, for that matter). My argument has used them extensively. If you want to understand how some 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. That's what methodological nominalism is banking on, at least insofar as it uses case studies as I have here. The standard approach, insofar as it uses case studies of scientific explanation, is banking on something more complex: the idea that looking carefully at examples of scientific explanation will be informative about *the properties 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 original questions, even when it is using the same case-study methodology. So if you want to spend

[24] Note that these are not the methods of experimental philosophy of science. That area of philosophy is firmly committed to the standard approach, using experimental methods to uncover what scientists think representation (or some other property) *is*, or what things must instantiate that property, not the role of the notion of representation in scientific explanation more broadly (see Favela and Machery 2022).

time elaborating an account of the property of representation, you need some reason to think it's going to be helpful — some reason to think that an account of a property, REPRESENTATION, will tell us more about *how and why representational explanation works* than an account that is, like the one I've given, explicitly and directly about *how and why representational explanation works*. Otherwise what would the theory of neural representation, and all the complication involved in its detour through the metaphysics of representation, be for? This is a challenge that proponents of the standard approach have not answered, and, to my knowledge at least, have not even been pressed to answer.

To summarize, methodological nominalism has the important advantages of offering a wider range of methods for philosophy of neuroscience, and of re-framing our current methods to answer our questions more directly. The second kind of advantage I want to discuss 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, if 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 a result of their conflict with Ptolemaic theories; the source of that anti-realism's significance was that it weighed in on a genuine scientific debate (Hacking 1983, 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 is generally taken to be *relevant to neuroscience itself*, and 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 whether something meets the criteria to *be* a representation. They are debates about models and explanations — which are more predictive, simpler, accurate, and so on.

E.g., consider how Shenoy et al. (2013), along with the rest of the motor control community (exemplified in Wang et al. 2022), understand the debate between representational and antirepresentational — specifically, *dynamicist* — approaches to motor cortex. It is not a debate about a property, REPRESENTATION, and whether motor cortex instantiates it. It is a debate about whether to model motor cortex as controlling motor activity through operations over neurons tuned to particular environmental and bodily variables, or as "generat[ing] motor commands by autonomous temporal evolution" (Wang et al. 2022, 796). It is a debate over whether we should model motor cortex in terms of the various environmental and bodily variables we can think of it as representing, or in terms of dynamical equations that describe its evolution through a state space given various starting-points (see Favela 2021, for a more thorough description of the debate). The anti-representationalism in that debate is a long way from the anti-representationalism typical in philosophy, exemplified by Chomsky's eliminativism (Chomsky 1995) and Hutto & Myin's dynamicism (Hutto and Myin 2014). Those views target the property of representation and some supposed incoherence or difficulty within it, and on those grounds reject the representational approach. On the view I've defended, the representational approach does not rely on any property of representation, and the debate between representationalism and anti-representationalism in philosophy can be understood as precisely the same debate as the one in neuroscience, from a slightly different perspective. It is a debate over the right explanatory stance to take on some capacity or brain area, where the right explanatory stance is determined not by whether the brain instantiates the property of representation, but by whether representational notions and the resources they introduce generate models that predict and generalize well, connect to their explananda, and so on. And it's worth noting that on the specific account I proposed in the last section, representational explanation and its strategies look just like what representationalists and anti-representationalists are arguing over in motor cortex: correlating brain activity with salient environmental variables to discover structures that model that activity.

So in addition to providing a satisfying account of representational explanation, and providing a wide range of methods that are well-suited to our goals as philosophers of neuroscience, methodological nominalism also puts philosophers in a position to join genuine neuroscientific debates, in a way that the standard approach, and its questions about how to define the property of representation, do not. I want to move on now, and address the skeptic. For lack of space I'm going to be very selective about the objections I discuss. I'll focus on the ones I come across the most, which don't concern the details of my account of representational explanation, but rather the plausibility of methodological nominalism more broadly. Hopefully by dispelling these objections I can clear the way for deeper and more informed ones.

The first objection is a superficial but quite tempting point: neuroscientists say the brain represents such and such, ask where its representations are, and so on (cf. Bechtel 2016; Ramsey 2021). Doesn't that mean they talk about representations? This is a sensible objection to *entity* anti-realism about representations. But the methodological nominalist has no objection to talking about the relevant entities. Nor does she want to avoid characterizing them in representational terms; she just objects to the idea that by characterizing them in representational terms we're making metaphysical commitments about their nature or the properties they instantiate. Bechtel's broader point in that paper, that representational notions play a role in experimentation (especially characterizing targets of discovery and manipulation) as well as model-building, can be fully accommodated on my view. In fact, it is a natural prediction of the account I've given: how could the terms in which we characterize explanatorily important parts of a system's causal structure and connect its behavior to explanatorily important environmental variables *not* figure into experimentation, help describe our targets for discovery, or informatively characterize the entities we manipulate? In light of this, Bechtel (2016), and this sort of concern in general, usefully complement the account I've given here. They certainly don't give us a reason to go on from entity-realism to try defining a property that the notion of representation may refer to.

Another concern is captured by Burge, in response to Dennett's instrumentalism: "Science invokes representation as a kind embedded in law-like patterns. So there is empirical reason to take it as a real kind in the world" (Burge 2010, p. 3). I've distanced myself Dennett's instrumentalism, but the objection applies to any view that doesn't take the standard approach's Step 2: science invokes representation as a kind, so there is empirical reason to take it as a real kind in the world — *and*, more to the point, empirical reason to concern ourselves with defining that kind when we try to understand representational explanation. How does this play out as an objection to methodological nominalism specifically? Well, in cognitive science we are confronted with patterns of behavior that we describe and explain in representational terms. The objection says that the existence of these descriptions and explanations is evidence for the existence of a property of representation, and reason for us to try defining it. But, of course, if the account in the previous section is plausible, it is a counter-example to precisely this claim. What cognitive scientific practice licences is a belief in the reality of what's modeled, the *stuff* of the brain, not in the reality of any properties that certain folk-cum-technical notions used in the modeling process might refer to, and much less in the need to define those properties to understand cognitive scientific explanation. But the objection is still useful, because it points the way to better ones. A substantive objection could either target the specific account I've given of representational explanation (i.e., reject it as a counter-example to Burge's claim), or it could show that something specific about cognitive science supports the inference Burge wants to make: an inference from scientific practice to the necessity of the standard approach. There may be interesting objections to hang on that scaffold, but, while I wait for them to be proposed and argued for, I'll move on.

Another common concern is about methodological nominalism run amok: if it works here, won't it work everywhere? And if it works everywhere, what happens to properties? Wouldn't it be troubling if every notion received this treatment, so that I was only (say) *guilty* in the sense that thinking of me as guilty introduces resources you can use to understand me? And even if we limit ourselves to scientific cases, wouldn't it be strange if we had to do science altogether without properties, if we couldn't understand claims in physics like, "The universe is made up of particles/fields/...," as introducing important categories needing definition, or as making metaphysical commitments?

In both the scientific and non-scientific cases, the answer is contained in the objections themselves: "Wouldn't it be troubling if ...;" "Wouldn't it be strange if ...." Good. The next step is to say what's troubling about it. If methodological nominalism about guilt is troubling, that is presumably because a theory of guilt-ascription that neglected the actual property, GUILT, would fail to meet some desideratum. It does seem to fall short at least of our desire for guilt to justify punishment — my being able to *fruitfully think of you in terms of guilt* isn't up to that task. So methodological nominalism is not so hard to undermine. This is true for scientific cases too. I do think methodological nominalism is plausible outside of representational explanation and outside of neuroscience. But where it isn't, it will be because there is some desideratum that it doesn't meet. It may be that physics has the ultimate goal not just to model the way the universe brings about states of affairs, but to taxonomize its basic constituents and say how they relate to each other and compose other constituents. Methodological nominalism will founder on that kind of case: the scientific project is explicitly to define properties, or to do something that clearly requires us to define properties. I don't want to commit to this understanding of physics. I just want to allay the concern that methodological nominalism will be hard to contain. It can be rejected anywhere there is some desideratum it fails to meet, and that kind of case might be fairly common.

And note (returning to a more superficial worry) that even where methodological nominalism is accepted, it doesn't keep us from talking about representations, protons, or whatever else. It *would* be strange for a view of (say) biology to tell us we can't talk about trees, or we can't say that fish exist. But a methodological nominalist about trees and fish would have no problem with either. Her problem is with a further step — the idea that this language calls for a careful definition of the properties TREE and FISH. She thinks we can account for everything we're really doing when we use the notions of *trees* and *fish* just by looking at the broader role of those notions. Trees and fish are particularly instructive cases because they form such motley, jumbled, and difficult-to-define categories that it is very common to hear biologists say there's no such thing as a fish (Banister and Dawes 2005)[25] or that it's impossible to say what a tree is (Ridley-Ellis 2019). That is, we should probably be skeptical of attempts to define these terms rigorously. And so we should be interested in other ways of understanding the function and obvious acceptability of claims to do with trees and fish, and explanations that advert to them, e.g., explanations of climate change or the

[25] This claim is often attributed to Stephen Jay Gould, but I haven't confirmed its provenance.

amount of algae in a pond as being due to the presence of trees and fish. One way of understanding methodological nominalism is as pointing out that these *other ways* of understanding concepts' roles in explanation are just as available when we're talking about concepts that might be susceptible to a definition (like *representation*) as they are when we're talking about concepts that probably aren't (like *trees* and *fish*). But, back to the present point, even if methodological nominalism *does* run amok, it's not really so troubling; it doesn't keep us from talking about protons, trees, or representations in pretty much exactly the way we already do. It undermines only a *further* philosophical, metaphysical, or definitional project — a project that we should already suspect is optional, given the cases of fish and trees.

The final concern I want to raise is about the relationship between my view and another, called pragmatism or deflationism (Egan 2019, 2021; Mollo 2020; Cao 2022). Deflationists accept the standard approach and the three-step tactic, and address themselves to the question, *what is neural representation?* But they answer with a deflationary definition or metaphysics: one that is distinctive in its sparseness and interest-relativity, and issues in a set of answers similar to those I gave in the first paragraph of this section. This is often expressed, by deflationism's main champion, as the idea that ascriptions of neural representation are just ways of "glossing" the *non-representational* characterizations of brain activity that constitute genuine scientific theories (Egan 2018). This gloss is supposed to serve various pragmatic purposes, but is not part of the scientific "theory proper" (Egan 2021, 41). Clearly, I've argued that representational notions do more than just *gloss* theories; they are involved in every stage of theory-construction and model-building, and are fully integrated into science and scientific theory proper. But I don't want to pursue this disagreement here. I want to bring out a more important difference between deflationism and methodological nominalism.

To bring out that difference, look at the similarities another way: you could arrive at a deflationist view by adopting my account, and simply adding what Richmond has called a "metaphysical appendix" (Richmond 2022): a definition to the effect that whatever is treated representationally in the way I've described *just is* a representation. You would end up with a deflationary, pragmatic, or otherwise “light” (Egan, in conversation) metaphysics of representation, in that the property of representation, and whether something instantiates it, is partly dependent on “explanatory practice in cognitive neuroscience” (Egan 2021, 43). This would allow you to hold on to the standard approach, and apparently without any additional or contentious commitments. But, though I consider myself to be in league with the deflationists, I think their addition of the metaphysical appendix is unhelpful. I’ve given you the account post-appendectomy, and I want to convince you not to open it up to shove the organ back in.

My main concern is that even if we can give a deflationary account of the property of representation, that property is not where the action is. The action is in the tools and strategies that representational notions introduce to scientific projects. A deflationist who acknowledges these tools and uses them to define representation is using the right resources, but *shifting the focus away from those resources*, back to the attempt to define representation. This is troubling since it was the turn away from defining representation that focused our attention on the new resources in the first place, and that made it possible to identify more roles for representational notions than just glossing ~~‘real’~~ theories. The shift back to definition is also troubling because it undermines the depth of the idea that motivates methodological nominalism, and that (I think) should motivate deflationism: the insignificance of a property of representation, deflated or not, to understanding representational explanation. Deflationism *does*, according to the methodological nominalist, constitute an advance on the question of representational explanation. But a methodological nominalist wants to take that advance farther, and to push it *deeper* into our understanding of explanation. Methodological nominalism is not just a new answer to the question, *what is neural representation?* It is a new approach to understanding representational explanations, not by defining their terms but by understanding their role in a broader explanatory economy. It is a new set of resources to appeal to, not definitions of properties but the range of facts about what explanations do and how they serve the myriad goals that make up the scientific project. And it is even a new range of methodologies: I’ve suggested that as long as we’re concerned with the *cognitive* role of representational explanation, we can bring the methods of experimental psychology to bear on the question. This is all captured explicitly by methodological nominalism, and at best implicitly by deflationism.

I think this also explains the strange dialectical position deflationism finds itself in. I am probably testing your patience already with the length of this article, so I won't make this case in detail. But I invite you to consider the kind of objections that plague deflationism. For now, just consider the objection that it collapses to another view, either anti-representationalism (Neander 2015; Hutto and Myin 2021) or a form of representational realism no different than the received view (Neander 2015; Ramsey 2021). The deflationist may be able to answer these objections, and similar ones, case-by-case. But for the methodological nominalist they fail for one principled reason: these objections, and many others (for quite a list, see Neander 2015; Ramsey 2021), object to the way the deflationist characterizes the *property* of representation. Either the property is "depreciated" and "diminished" Ramsey (2021, 74), in which case the deflationist seems indistinguishable from the anti-representationalist, who argues that the brain doesn't instantiate any non-trivial property of representation (Hutto and Myin 2014). Or the property is robust and non-trivial, in which case the received view can try to accommodate the special features the deflationist attributes to that property Ramsey (2021, 74-76). But, of course, this kind of objection can't even get off the ground if we are methodological nominalists. We cannot be mistaken for the received view because we don't disagree with its definition of representation *in subtle ways it can accommodate*; we disagree with the very approach that makes that definitional task part of their inquiry. And we cannot be mistaken for anti-representationalists because we are describing *how and why representational explanations work* — once we understand that form of explanation, it is open to us either to endorse it (representationalism) *or* reject it (anti-representationalism).

I don't want to be mistaken as arguing *against* deflationism in these last paragraphs. As I've indicated, the deflationist will have her own responses to these objections, and I haven't evaluated them. But I've tried to show that methodological nominalism offers a principled and deeper response: these kinds of objection mistake what is, or should be, at issue — not the property of neural representation, but the way representational notions serve neuroscientific explanation. So I only intend to extend the deflationist an invitation to methodological nominalism. From the methodological nominalist's point of view, deflationists have taken a significant step forward from traditional philosophical approaches to neuroscientific explanation. The invitation is to extend their step farther, to see it as making a more fundamental point, and to, with the methodological nominalist, embrace Step 2* as the explicit starting-point for understanding representational explanation. From that starting-point, I think the deflationist will have a more thorough hearing, and the philosophy of neuroscience in general will have a methodologically-sound basis for a philosophically-illuminating account of representational explanation — and one that is also an intervention into 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., Muller, T. H., Whittington, J. C., 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*

Burge, T. (2010), *Origins of Objectivity*, Oxford University Press, New York.

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

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.
URL: *http://dx.doi.org/10.1016/j.cell.2017.05.011*

Chemero, A. (2011), *Radical Embodied Cognitive Science*, MIT Press.

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

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

Dennett, D. C. (1988), ‘Precis of The Intentional Stance’, *Behavioral and Brain Sciences* **11**, 495–546.

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. 690–689.

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

Egan, F. (2018), A Deflationary Account of Mental Representation, *in* J. Smortchkova, K. Dolega and T. Schlicht, eds, ‘What are Mental Representations?’, Oxford University Press, p. forthcoming.

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. (2022), ‘The Untenable Status Quo: The Concept of Representation in the Neural and Psychological Sciences’, *ArXiv* .

Fodor, J. (1987), *Psychosemantics*, MIT Press.

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*

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**(September), 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’, **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.

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

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

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

Ramsey, W. M. (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. (2022), ‘How Computation Explains’, *ArXiv* .

Richmond, A. (n.d.), 'What Really Lives in the Swamp? A New Monster for Etiologists', *Forthcoming* .

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

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.

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.

van Fraassen, B. C. (1980), *The Scientific Image*, Oxford University Press.

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*