Methods in Mind: Explanation in Cognitive Science

Andrew Richmond

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Methods in Mind: Explanation in Cognitive Science

Andrew Richmond

Together, these three papers aim to develop scientifically informed accounts of the role of computation and representation in cognitive science. Along the way, they illustrate and defend a *methodologically nominalist* approach to the philosophy of cognitive science: one that investigates scientific explanation by setting aside any properties that scientific concepts might refer to, focusing instead on the concepts themselves and their role in cognitive science's explanatory economy — what they help scientists to explain, and how. In addition to these philosophical upshots, the papers intervene on a number of debates within cognitive science itself.

Chapter 1, "How Computation Explains," discusses the monumental shift in our understanding of the brain triggered by the project of computational cognitive science: the use of tools, concepts, and strategies from the computer sciences to investigate the brain. Philosophers have typically understood this project, and the *computational explanations* it provides, to assume that the brain is a computer in some sense — a sense to be specified by the metaphysics of computation. That metaphysics, by revealing what exactly we attribute to the brain when we say it computes, is supposed to show how and why computational explanations work, and in doing so to provide a philosophical foundation for them. In contrast, I give an account of computational explanation that focuses on the resources computational explanations bring to bear on the study of the brain. I argue that computational explanations help cognitive scientists build perspicuous models that capture precisely the kinds of causal structures they seek, and that no metaphysics of computation

is required to understand how they do this.

Chapter 2, "Computational Externalism," argues that the brain does not have a computational structure in itself, but only in conjunction with things outside of it. I understand this thesis in one of its more extreme forms, and one more in line with what the previous chapter says about computational explanation. The externalism I defend does not just say that the computational identity of a system depends on its environment — that would be the case if, e.g., a system's computational identity depended on the representations it processed, and the identity of those representations depended on the environment. Instead, it is *the brain's causal structure, insofar as computational explanations aim to capture it*, that the brain does not have on its own. I support this with a case study in the neuroscience and evolutionary biology of color vision, which poses a challenge for internalists about computational structure. And I go on to draw out some conclusions of computational externalism for debates within cognitive science.

Chapter 3, "What is a Theory of Neural Representation For?", 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" is acknowledged to mean "representation as cognitive science uses that notion." But eliding this clause allows philosophers to focus more squarely on the definition, or metaphysics, of neural representation. 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 — and not at all important to understand what neural representation itself is. I propose that representational notions help us construct and understand models of the brain's causal structure, and support principled idealizations of that causal structure. I describe the way that representational notions serve these purposes, focusing on the way they support the imaginative projection of structures from one domain (the

'represented') onto another (the 'representing') as models. I show that understanding these functions does not require us to define the property of representation, or even commit to its existence. It does not require us to reject its existence either, though: we need not be anti-realists but merely methodological nominalists, focusing on how technical notions help scientists achieve their goals, and not (at least without good reason) on the nature of any properties those notions might refer to.

Table of Contents

Acknowledgments iii

Dedication vi

Preface 1

Chapter 1: How Computation Explains 6

1.1 Introduction 6

1.2 Triviality 9

1.3 Domain transfer and a pragmatic approach to computational explanation 13

1.3.1 Making room for the pragmatic approach 13

1.3.2 How computational explanation serves cognitive science 15

1.4 Objections 24

1.4.1 Empirical content: another triviality problem? 24

1.4.2 (Yet) another triviality problem 26

1.4.3 One last triviality problem 27

1.4.4 Metaphysical appendices, naturalism, and objectivity 29

1.5 Conclusion 31

Chapter 2: Computational Externalism 34

2.1 Introduction 34
2.2 What is computational externalism? 34
2.2.1 Computational structure and computational explanation 34
2.2.2 Internalism and externalism 37
2.2.3 How could computational externalism be true? 41
2.3 The argument for externalism 45
2.3.1 Case study: color vision 45
2.3.2 Finally, the challenge 56
2.4 Externalist methodology in cognitive science 61

Chapter 3: What is a Theory of Neural Representation For? 66
3.1 Introduction 66
3.2 Examples: place cells and the fusiform face area 67
3.3 Questions about representation 69
3.4 Methodological nominalism 73
3.5 An account of representational explanation 75
3.6 Upshots, objections, and miscellany 88

Coda 98

References 108

Acknowledgements

I have no idea where this dissertation came from. I had planned to write about Kant, Frege, and formal epistemology (don't ask. . .), and then suddenly I was sitting in a lecture on retinal ganglion cells, working out a paper on the way we individuate parts of the brain. (The thesis, and forever the dissertation topic that got away: *there are no brains*.) Anyways, I stumbled from Kant, to the retina, and then, by another series of accidents, to the dissertation you hold right now. I can only apologize.

But, although I'm not much for sharing credit, I'm happy to spread the blame. Insofar as this dissertation reflects a faith in the open horizons of philosophical progress, Jenann Ismael is at fault. She, more than any other philosopher, showed me how to see which questions really matter, and how to set aside the rest. And it was in her work that I began to see the progress you could make without deferring to disciplinary boundaries and traditional conceptions of those questions.

To the extent that the dissertation gets anything right about the deep and enduring questions that shape our philosophical projects, Chris Peacocke is to blame. At a number of points in this dissertation (most importantly, when I thought I was becoming a naturalist), Chris's feedback has cut straight through to the heart of the issue I was struggling with and changed the direction of my thinking.

And if it comes across that this dissertation has been informed by hundreds of patient conversations, insightful suggestions, and the painstaking excavation it takes to turn poorly-formed intuitions into tight arguments and carefully-informed exposition, for that John Morrison must be held accountable. When I read through this one last time, I can hardly find a paragraph that doesn't

include an idea, framing, emphasis, or objection that came out of our discussions. His feedback and guidance have, more than anything else, shaped this dissertation and the philosopher I've become. To whatever extent it looks like this dissertation was written by someone who knows how to do philosophy, it's John's fault for teaching me.

At many stages of the dissertation Rosa Cao and Frankie Egan provided a sympathetic ear and detailed, incisive feedback. Because my approach is so close to their own I tend to emphasize differences, but Rosa and Frankie are among the pioneers of the kind of work I'm trying to do. Their support has been invaluable, and far more generous than I probably deserve. Unfortunately, since they are not employed at Columbia they had no professional obligation to offer so much support — they did so purely out of their own kindness and generosity — and their share of the blame must be increased accordingly.

For many, many hangovers, I blame Yarran Hominh, Billy McCarthy, and Phil Yaure. Also for the hangover-preceding conversations, which managed to be not only coherent, but some of the most illuminating discussions I've had about this work. (I won't comment on what that says about the work...) For my sense of camaraderie and my certainty that I have a place in this discipline, the guilt belongs to those three along with Noah Betz-Richman, Samara Burns, Nemira Gassiunas, Natalie Hannan, Michael Holmes, AJ Marsh, Jorge Morales, Devin Morse, Kate Pendoley, Stephan Pohl, the members of UAW Local 2110, Caitlin DeClercq and the rest of Columbia's Center for Teaching and Learning, and my students over the past five years.

For many headaches about big pictures, and for the clarity that (sometimes) resulted, I place the blame squarely on Amogh Sahu. For providing helpful comments, arguments, formulations, emphases, and bits of encouragement I charge David Barack, Simon Brown, Mazviita Chirimuuta, Andre Curtis-Trudel, Joe Dewhurst, Raphael Gerraty, Linus Huang, David Rosenthal, Achille Varzi, Jason Winning, the members of PopRocks, the attendees at the Columbia Grad Student Workshop, and the audiences at countless philosophy and neuroscience conferences.

I would never have finished this dissertation without the expertise, advice, and especially the patience of Maia Bernstein, D.D. Meakin, Asha Panduranga, Stacey Quartaro, and Clay Rains, so,

unfortunately, they are also accessories to the dissertation that has resulted.

It is almost entirely the fault of Sara Cummings that that I made it out of grad school with my health intact and with a fulfilling non-academic life. And for the fact that this dissertation exists at all, you'll have to blame Wayne and Marg Richmond, who taught me how to commit to an accident and make something meaningful out of it. This dissertation wouldn't exist without all these people, the communities they built, and the person, teacher, and philosopher they helped me become.

Oh, and that time I shouted "fuck" in a Zoom talk was Joanna's fault.

Dedication

For Charles Manson

(*Not that one*)

For the carpenter's and the geometer's inquiries about the right angle are different also; the carpenter restricts himself to what helps his work, but the geometer inquires into what, or what sort of thing, the right angle is, since he studies the truth. We must do the same, then, in other areas too, seeking the proper degree of exactness, so that digressions do not overwhelm our main task.

— Aristotle, *Nicomachean Ethics*

Preface

This is a short dissertation — a veteran philosopher will hardly break a sweat over three papers. But I want to make some preparatory remarks anyways. Specifically, I want to talk about the context and goals of these three papers.

I am motivated, in the first place, by a simple problem. Cognition didn't used to make much sense. The brain, at least, didn't used to make sense at all. For a long time it was hardly conceivable that the brain could be the organ of cognition. And even when — rather late in the history of science — it became widely accepted that the brain is the seat of cognition, this view was clumsily supported. We had many (understandably) immature approaches to the brain, starting from a focus on ventricles (cognition is pumps pushing fluids around) and moving on to comparisons with telegraph machines, steam engines, and other unlikely contraptions (see Cobb, 2020, for a short overview of this history). Eventually we learned more about the fine-grained structure of the brain itself, most importantly settling on the neuron doctrine (G. M. Shepherd, 1991). When we started focusing on neurons, things moved a bit more quickly. And after the dark ages of behaviorism things really picked up, with an explosion of increasingly sophisticated theories about the neural mechanisms of cognition (again, see Cobb, 2020).

Now the brain makes a bit of sense. We don't know enough about the particulars of brain function, but we have a solid and maturing toolkit for investigating them, as well as some promising models of particular aspects of cognition. You can pick your favorite example here — color vision, spatial navigation, sound localization, whatever. The simple problem I teased above is this: *what happened*? Why does the brain make sense now? What are we doing *right*?

Let me set aside two related questions. First, does the brain *actually* make sense? Or does the BOLD signal deceive? Are we misled by our neglect of glial cells? Are we radically wrong about the brain in these and other ways, setting up future historians of science to look back on us like we look back on the caloric theory of heat? In all modesty, this is probably true. Few scientific approaches survive scrutiny if you scrute them from far enough in the future. But if Whig history is a vice, so is the detached cynicism that takes this 'pessimistic meta-induction' too seriously. We don't need our current science to work in the far future; we need it to work now, and to work better as time goes on. And we don't need it to work better *sub specie aeternitatis*, but better by our own standards — the only ones we have any access to. So what I'm concerned with is the practice of cognitive science today: the way it works, makes progress, reasons, explains, and understands. We make sense of the brain a certain way, and, by our own lights at least, we do that a lot better today than we did fifty years ago or a hundred. What exactly is cognitive science doing, then, and how should we understand it?

The second question I want to set aside is the historical question of what happened. That is not a project for a dissertation; it is a project for a two-volume, 1400-page manuscript already written by Margaret Boden (2006). I won't add anything to the history of cognitive science. My starting-point is that something is special about current cognitive science. A historian would put the contrastive emphasis on *current*: What is cognitive science doing now that at other times it was not? I put the emphasis instead on *special*: What is cognitive science doing, and perhaps doing well, that other approaches might not do?

I approach this question with a few background commitments. The first is a commitment to naturalism, in a broad sense. I don't mean a commitment to natural kinds, a deference to scientific knowledge, or the view that the scientific image is all that really exists. I mean naturalism *about* science, in Dewey's sense: a commitment to understanding science as a human production, a result of complex social, institutional, and especially practical factors, not as divining eternal truths but as solving the problems that we set for it, or that it sets for itself as it takes on a life of its own. Nagel's summary of Dewey's pragmatism in logic sums up this commitment well:

> Just as plants and animals are studied to the best advantage only within their natural environment so that the distinguishing traits and uses of their various organs may be ascertained, so on Dewey's view the specific techniques and principles of logic receive an adequate theoretical interpretation only by being exhibited in the roles they play within the process of inquiry. (Nagel, 2008, pp. iix–ix)

I also come at this work with a commitment to interdisciplinarity. John Tukey used to say "the best thing about being a statistician ... is that you get to play in everyone's backyard" (Brillinger, 2014). Many philosophers have a similar attitude. E.g., it's common to hear stories about people going to grad school in philosophy because it let them study whatever they wanted: they could spend their time learning and thinking about any subject, X — they just had to report back to their advisor that they had been "working on the philosophy of X." But it's hard to jump into a game you don't know well, and some philosophers prefer to look over the fence at (say) biologists playing with the word *function*, make themselves a quick sketch of the game, and come back to discuss it exclusively with other philosophers (perhaps occasionally peering back over the fence to confirm some observation about the game, or worrying about whether they've characterized the game correctly). In contrast, I take the ideal to be engagement with cognitive scientific concerns *themselves*. I don't want to take up the notion of representation as it is used in neuroscience and come back to have a conversation about it with philosophers, focusing on philosophers' concerns and preoccupations, e.g., about the nature of the representation relation. I take my audience to consist of cognitive scientists as well as philosophers, and I want my work to reflect the concerns and needs the two disciplines share. So each of these papers intervenes on a debate of philosophical significance, but one that is also playing out in cognitive science journals. For obvious reasons, my resources are largely philosophical, but (and this is a central theme of the papers) I leave aside philosophical questions or debates that don't promise to serve interdisciplinary goals — ones that would have me peering over the fence rather than joining in the game, at least in some minor role. It's not that philosophers have no worthwhile games to play themselves, but philosophers of cognitive science tend to claim that their work is not just *about* cognitive science but *relevant* to it.

A main goal of this dissertation is to take that ambition seriously — to draw peering philosophers over the fence, even if it means they have to leave some of their own games behind for the time being.

Finally, and because of the previous two commitments, I am committed to an approach that some might consider 'sociological,' the kind of approach exemplified by, e.g., P. F. Strawson and Cora Diamond. I don't think there's a sharp dividing line between this approach and other, more traditional ones. Strawson wasn't taking up a different question than other philosophers when he suggested that we approach questions about moral responsibility from within the web of beliefs and concerns that condition our practices of holding people responsible (Strawson, 1974). Diamond wasn't changing the subject or stepping outside of academic philosophy when she suggested that animal rights had to be justified from within our everyday practices and practical understanding, particularly the aspect of that understanding that draws a significant, even if 'ultimately' unjustified, moral line between human suffering and animal suffering (Diamond, 1978). What both aimed to do was to clarify the purpose of their philosophical subject in terms of the concrete problems it was supposed to bear on — the tensions or problems that give rise to philosophical questions, and which answers to those questions should help us resolve. Philosophers are easily swept up in side-problems and easily tempted by sub-questions, especially ones about definitions or natures. E.g., we want to understand cognition, so of course we spend time thinking about its defining characteristics. But this raises the question of how exactly cognition is to be defined, and we generate new and more complex definitions of cognition until this takes on life as a project of its own.[1] But defining cognition is patently not the same as understanding cognition, and the former is not even a plausible way of making progress on the latter. Questions about cognitive science run the same risk, particularly since they often concern cognitive scientists' use of contentious concepts like *representation*, *computation*, and *function*. We want to understand the explanations these concepts figure into, but this question sits alongside some highly distracting ones about the definitions or natures of representation, computation, and function. Like Strawson and Diamond, I want to

[1] See Allen (2017) for a discussion, and a convincing rejection, of this project.

avoid those distractions, and focus instead on the real and practical challenges that give rise to our philosophical questions — the challenges that philosophical answers should, ultimately, help us overcome.

This all means that I want to address questions that are not just about science but that are of genuine importance *for* science, and I want to address them with careful attention to the problems and constraints that make them important for both philosophy and science. I want to play in cognitive science's backyard, without losing the characteristic reflective attitude of philosophy, but reflecting and philosophizing from within the constraints of the concrete, practical projects that constitute cognitive science.

That is perhaps an overly doctrinaire note to begin on, and certainly too optimistic — it will be the work of a career, not a dissertation, to live up to these ideals. But this is nonetheless the background for these three papers, and the context that, I hope, will make their contributions clear.

Chapter 1: How Computation Explains

1.1 Introduction

Cognitive science gives computational explanations of behavior. From neuroscience in particular we learn that the brain sees depth by computing the disparity between retinal images (Nityananda & Read, 2017), discriminates colors using cone-opponent computations (Thoreson & Dacey, 2019), localizes sounds by computing inter-aural time differences (Grothe et al., 2010), and supports reaching and grasping tasks by computing vector displacements (Shadmehr & Wise, 2005). Cognitive scientists also make general appeals to the computational capacity of neural channels (Gallistel & King, 2009), the computational architecture of the brain (Lake et al., 2017; Yamins & DiCarlo, 2016), and the types of computation it is capable of (Danks, 2019).[1]

Perhaps this has all become commonplace enough to dull our critical instincts, but if we dwell for a moment on this fact, this explosion of computational explanations over the past half-century, some questions arise. What *are* these explanations? How do they work? What distinguishes them from other kinds of explanation? And why have they been so successful? (Not just successful, but so successful that it is hard to imagine cognitive science without them.) In its simplest form, the question is: *how* and *why* do computational explanations — explanations that characterize the brain in computational terms — work?

The received view is that computational explanations work because the brain *is* a computer. The brain supports depth perception, e.g., by *literally computing* retinal disparity. If you accept this view, you will want to know what exactly it means — what it is to be a computer, and why so many successful explanations latch onto this property of the brain. So your inquiry has the

[1] Discussions of the computational approach in general (as opposed to specific computational explanations) are also common (R. Cao, 2019; Chirimuuta, 2019; Fodor, 1975; Gallistel & King, 2009; Hardcastle, 1996; Pylyshyn, 1984, 1993).

form of a traditional metaphysical question. What is a computer? What features make something a computer? What criteria must something satisfy to be a computer? Call this the Metaphysical Approach to computational explanation.[2]

Note that the Metaphysical Approach doesn't want a theory of computation like the one Turing gave. That was a theory of computable functions, i.e. functions for which there exist effective methods or algorithms, and the nature of those algorithms. These are formal, abstract things, these functions and algorithms. In the metaphysics of computation, we're not concerned with *formal systems* themselves. We want to know what it takes for a *physical system* to compute, i.e. to *implement* one of those formal systems. There are algorithms for addition, and then there are the cash registers that implement them. The question here is about the cash registers (and other physical systems) — what does it take for a hunk of metal and plastic to implement an addition algorithm? What makes the cash register a computer, and what makes it the specific computer it is?

So, the Metaphysical Approach wants to identify the features of the brain that make it a computer, and that therefore make computational explanations of it appropriate and successful. But a satisfying metaphysics of computation is hard to come by. And when accounts of computation are unsatisfying, or when someone's metaphysics of computation suggests that any old rock (Putnam, 1991), pail of water (Lycan, 1981), or brick wall (Searle, 1992) computes, one is liable to hear some familiar refrains. The brain isn't a computer; that's "just a metaphor." Or debates about what computation is and whether the brain satisfies that definition are "just semantics." A recent paper by Richards and Lillicrap (2022) exemplifies both frustrated responses: they argue that when we say the brain is a computer we either mean it's somehow like a laptop, which is just a metaphor (and not a very useful one), or we mean it implements a universal Turing machine, which is literally true but uninformative, reflecting only a stipulation about how we use the phrase "is a computer."

But even if the brain-as-computer-metaphor *is* a metaphor, it isn't *just* a metaphor — it's one

[2]For representative examples, see Chalmers (2011), Piccinini (2015), and Shagrir (2022). Things are no different if one thinks that computational explanations work because the brain *computes*, but is not *a computer*. The questions that arise are perfectly analogous (What is it to ~~be a computer~~ compute?), and the distinction won't be important for my purposes.

of the most successful explanatory approaches in recent science. That needs explaining, and "it's just a metaphor" hardly even begins to offer an explanation. And even if they are semantic in nature, debates over what computation is aren't *just* semantics. As I've described, they are part of a broader project that aims to show how and why computational explanation works. "That's just how we use the word" is barely a first step in that project, and if elaborated it will run into the problems I note in section 1.2. We can dismiss the semantic debates if we like — in fact, I do. But unlike my fellow travelers, I take it that I'm not *just* setting aside those debates. I'm also taking up a burden: to explain the success of computational explanation in some other way. This paper is my attempt to discharge that burden. I think all the resources I need can be found in less controversial domains than the metaphysics of computation: I will appeal only to the kind of things computational notions allow us to do as we investigate and especially model the brain.

I'll illustrate the metaphysics of computation by introducing the *triviality problem* in section 1.2. In section 1.3 I'll turn away from the metaphysics of computation, and instead treat computational explanation as an example of the more general phenomenon of *domain transfer*: the use of tools, strategies, or concepts in a novel domain, or for a purpose they weren't originally developed for — like when companies apply NASA's failure-detection strategies to ad campaigns rather than rocket components (Edsel, 2016), or when teachers use techniques from game design to make their courses more engaging (Miller, 2014). I'll give an account of computational explanation that appeals only to *the resources it introduces into cognitive science*, and show that the metaphysics of computation, whatever it may be, is irrelevant to this account: as far as computational explanation is concerned, there might as well be no such thing as a computer. Call this the Pragmatic Approach. I'll develop this approach in response to some objections in section 1.4, and conclude in section 1.5 with some final considerations that support it against the Metaphysical Approach. This will not just set the stage for further development of the Pragmatic Approach, but also clarify the terms of engagement with the Metaphysical Approach — the kind of argument for a metaphysics of computation that could, in principle, be compelling.

1.2 Triviality

In the examples I began with, the brain is not merely modeled computationally, like the weather often is (Ham et al., 2019). Weather models predict the future behavior of the weather, but often not the internal processes that bring it about. Genuine computational explanations, at least as they figure into cognitive science, do more than predict behavior. They are *process models*. They are supposed to explain a subject's capacities by telling us about the processes in the subject's brain that bring them about.[3] And, according to the Metaphysical Approach, computational explanations tell us that the brain brings those capacities about *by computing*, or by *being a computer*.

The burden is to say what exactly this means. What is it to be a computer? What criteria do we apply to tell whether something is computing, or what it is computing? The main hurdle for philosophers answering these questions is the *triviality problem*. Many seemingly plausible criteria answers to these questions end up counting too many systems as computers, and counting any given system as computing too many things. And on the Metaphysical Approach, this spells disaster for computational explanation. This section will describe the triviality problem in more detail, as a way of illustrating the Metaphysical Approach and its differences from the Pragmatic Approach.

It is standard to introduce triviality using the simple mapping account of computation (Egan, 2014), according to which a system implements a computation if its states mirror the stages of the computation, i.e. if there is a mapping between the stages of the computation and the system's states. On this account your calculator computes addition because when you punch "5" and then "7", the display shows you "12," and in doing so it has transitioned from states that map to the numbers 5 and 7 to a state (the output) that maps to the number 12. To compute some more detailed algorithm, a system must only transition between physical states in a way that preserves a mapping to the algorithm's more numerous stages. We can set aside some niceties of definition and

[3]This is not the only use to which computational models are put. They can be more than predictive models, but still less than process models. E.g., they can specify the optimal functioning of a system (Sánchez, n.d.), or explain why it is the way it is (Chirimuuta, 2014). The point is well taken, but these uses of computational explanation are not my target here.

say that *a system performs a computation just whenever its physical dynamics map to the dynamics of the computation*.

This is intuitive. Ask a computer scientist what makes something a computer and they'll likely give you the mapping account. But it has a problem: mappings are cheap. Virtually every system maps to virtually every algorithm or computation, given a suitable 'carving up' of the system in question. For instance, if we want a rock to compute the addition function, we need only decide which instances of addition we'd like it to perform in a span of time. If we'd like it to have just now computed the addition function for inputs 5 and 7, we take the past three seconds of the rock's existence and map its state at each second to one of the numbers: it transitioned from state-at-second-1 (mapped to 5) and state-at-second-2 (mapped to 7) to state-at-second-3 (mapped to 12).[4] There is nothing special about the rock. This is true of every object that has been in at least three states over the past three seconds. And it's not limited to simple computations. If we want the rock to compute the addition function using, say, the same algorithm your calculator does, we map its physical states over a span of time to the stages of the algorithm your calculator follows.[5] According to the simple mapping account, this would show that the rock computes addition just as your calculator does. This is treated more rigorously by Putnam (1991) and Chalmers (2011), but the upshot is simple: mapping relations are too numerous to constitute a metaphysics of computation, because they make it too easy for a physical system to be a computer. And that saps or renders mysterious the explanatory force of computation in at least two ways. I'll dwell on them for a moment, because they nicely illustrate the contrast between the Metaphysical Approach and my own.

First, much debate in cognitive science is over which computations the brain performs.[6] It is because the brain performs cone-opponent computations and not simple cone summations that it

[4]Repeating states — e.g., if the rock's next calculation includes a 5 as well — are handled by disjunctions, so it is state-at-second-1-or-4 that gets mapped to the number 5.

[5]And of course if we carve up the rock's states more finely, we can map it to complex computational structures like whole Turing machines or neural networks.

[6]We might say the brain is performing every computation, but only some are explanatorily relevant. But this just pushes the question back a step: we have to say what makes a computation explanatorily relevant, and that framing doesn't seem to offer any additional traction on the issue at hand.

supports the kind of color vision it does (Jacobs, 2014). If it performed both computations, the connection between both models and their explananda would be severed: cone-opponent models couldn't make a prediction about color vision that cone-summation models didn't also make, and vice versa, because each would have to allow that the brain also performed the other model's computations. It is because the brain performs certain computations *and not others* that those computations explain its capacities. Consider also the *discovery* of the brain's computational properties, which the mapping account renders far easier (just find a mapping — as easy to do with the brain as with a rock) than the history of cognitive science would suggest.

And second, consider systems other than the brain. Even if we find sufficiently narrow criteria so that the brain computes only a limited set of functions or algorithms, it is a problem if too many other things also compute them. If it turns out that a rock implements an addition algorithm, then your brain's implementing that same algorithm could not explain its arithmetical performance, because that algorithm can be implemented without supporting arithmetic. Performing a computation needn't be *sufficient* for addition — other background features may be involved. But if those background features are computational (e.g., the computations the brain performs to use the outputs of the arithmetic module), the same problem arises: the rock will have them too, on the simple mapping view. And if the background features are not computational, then non-computational properties do all the work making the difference between a system capable of arithmetic and a system incapable of it; computation is irrelevant, and again we've undermined its explanatory role.

The problem, then, is that rather than illuminating computational explanation, or providing it with philosophical foundations, the mapping account appears to *undermine* computational explanation and make its success mysterious. If we're approaching computational explanation through the metaphysics of computation, these problems have to be solved by an appropriately narrow definition of computation: one that limits which systems implement which computations, in a way that preserves the scientific role of the notion of computation. E.g., we might consider the causal view (Chalmers, 1996), which requires an algorithm to map to a certain kind of causal structure in a system. Other attempts have grounded computation in not just the causal but the teleological

(Milkowski, 2013; Piccinini, 2015) or representational (Peacocke, 1994; Shagrir, 2018) properties of computing systems.

It hardly needs to be said that these views are all controversial. It is debatable whether the causal view alone solves our problem. Scheutz (2012) and Shagrir (2001) argue that the causal view still allows a problematic proliferation of computations, and Egan (2012) argues convincingly that even if it is safe from triviality, Chalmers' definition of computation is not suitable for the use cognitive scientists put the notion to.[7] And teleological properties — properties to do with something's *function* or *purpose* — incur severe explanatory debts themselves, so it's not clear they put the metaphysics of computation on better footing (cf. Dewhurst, 2018).[8] And of course there are similar, if less severe challenges for most views of representation too (e.g., see Egan, 2019).[9] I note these issues not as an argument against the causal, representational, or teleological views of computation, but only to make clear the problem that the Metaphysical Approach revolves around: defining physical computation so as to include all the right systems and exclude all the wrong ones. This is supposed, on the Metaphysical Approach, to be the first step in explaining how computational explanation works.

What is distinctive about the Pragmatic Approach is that the triviality problem and all of the resulting puzzles will be irrelevant to it. The main desideratum for the Pragmatic Approach is the same as for the Metaphysical Approach: to explain how and why computational explanation works. But the Pragmatic Approach does not achieve this via the metaphysics of computation — instead, it looks to the goals computational explanation serves and how it serves them. If this approach successfully explains computational explanation, it will have shown that the metaphysics of computation is unnecessary: a careful look at how computational notions work, and what they do for cognitive scientific explanations, is enough. This is what I'll try to show in the next section.

[7]The causal view also appears unable to account for the apparent environmental individuation of computational processes (Richmond, n.d.-a; Shagrir, 2001; Shea, 2013).

[8]See Chalmers (2011, p. 334) and Sprevak (2019, p. 177), on the need for computation to be grounded in well-understood notions.

[9]Doubly so for the most popular view of representation, which grounds it in teleological properties (e.g., Neander, 2017). But on my own view, representation is an important part of the story, and representationalists about computation will find that I preserve much of what is most important about their view (Richmond, n.d.-c).

1.3 Domain transfer and a pragmatic approach to computational explanation

The goal of this section is to build an account of computational explanation in the space between two extremes. First the account shouldn't pull us into debates about the metaphysics of computation, or about whether the brain *really is* a computer. But, second, it should also explain the ubiquity and success of explanations that conceive of the brain in computational terms — it should say more than "it's just a metaphor."

1.3.1 Making room for the pragmatic approach

First, let me reiterate something from the previous section. I distinguished between merely predictive computational models and genuine computational explanations. One way of making this distinction more precise is to say that computational explanations give *process models* of their target systems — models of the processes that generate their behavior. Simon and Newell expressed this early on:

> We do not say that we understand the magic [trick] because we can predict that a rabbit will emerge from the hat when the magician reaches into it. We want to know how it was done — how the rabbit got there. Programs like LT [the authors' "Logic Theorist"] are explanations of human problem-solving behavior *only to the extent that the processes they use to discover solutions are the same as the human processes*.[10] (Simon & Newell, 1973, 147, my italics)

So there is the distinction between merely predictive models, like computational models of the weather, and process models, which detail the processes a system undergoes to generate its outputs. But process models are not necessarily models that attribute, to their target systems, membership in a special category. Consider models of physical bodies expressed in calculus equations. These models can be more than predictive — they can model the processes that bodies go through to

[10] Marr (1982, p. 23) expresses a similar sentiment. Also see Fodor (1968), Kriegeskorte and Douglas (2018), R. Sun (2008), and Nancy Dice in Bailer-Jones (2002) on computational models as process models.

generate their dynamics or final positions. But though the model is couched in calculus, the system it models need not be a *calculizer*, or meet some criteria for membership in that category. We don't think the model unveils any *inherent calculus-properties* of the system. Thinking that way about *computational* process models in particular would be a sharp divergence from our treatment of most scientific modeling. For an even starker example, consider models of fluid dynamics applied to traffic jams (D. Sun et al., 2011) or epidemiological models of disinformation (Kucharski, 2016). These models don't require or assume that traffic literally is a fluid, or that disinformation is, by some criteria, a virus.

These are just examples of domain transfer: tools, concepts, or strategies that were developed for one purpose or domain are being applied to another. Hospital teams have borrowed strategies from Formula 1 pit crews and dance choreographers to hand off patients from surgery to the ICU (Sower et al., 2008). They are not performing a ballet, and their patients are not assumed to meet the criteria for the category FERRARI or MCLAREN. Nor do we need a metaphysics of race-cars to understand how and why the hospitals' strategies work. All we need is to understand how the tools from one domain work, and why they work in another. And, at the risk of belaboring the point, these questions need not be answered by saying that the elements of the two domains share a status as viruses, fluids, or race-cars.[11] So it is premature to say, as Chalmers does, that "[w]e cannot justify the foundational role of computation [in cognitive science] without first answering the question: *What are the conditions under which a physical system implements a given computation?*" (Chalmers, 2011, p. 325). We'll know whether we need to answer that question only once we understand what we're doing when we give computational explanations, and how this approach serves cognitive science. If the approach works like a typical domain transfer, the project Chalmers describes is unnecessary. In section 1.3.2 I'll argue that computational explanation *does* work like a typical domain transfer.

[11] They must share something for the domain transfer to be successful — something about the two domains must explain why the same tools can be used in both cases. The point is that it would be silly to think this something was their status as race-cars, rather than that pit crews working on a car do so with certain goals and under certain constraints, and that the goals and constraints for the hospital team are comparable in some respect.

1.3.2 How computational explanation serves cognitive science

Cognitive science wants to explain the cognitive capacities of complex systems like biological organisms: the capacity to detect and distinguish between stimuli, to decide on a course of action, to navigate spatial and social environments, and so on.[12] These are capacities to produce appropriate behavior in a range of environments and given a range of inputs. And cognitive science, since the rejection of behaviorism, aims to explain these capacities by appeal to the internal causal structure of the system in question. It is this structure that mediates the system's behavior; it is this structure that determines its solutions to the problems it faces; and it is this structure in terms of which we understand that behavior and those solutions.[13]

These goals call for a description, at an appropriate level of detail, of the brain's causal organization and processes, along with conceptual resources to *explain* behavior under that description. This means we need, at least:

(A) A language or formalism with which to describe the causal structures in the brain that support cognitive capacities.

(B) Conceptual resources with which to form questions, hypotheses, and explanations regarding those causal structures and the way they support cognitive capacities.

(C) Heuristics and background knowledge that make it efficient to form and work with these hypotheses and explanations.

The need for expressive and detailed languages and formalisms, as in (A), is widely discussed. See Lazebnik (2004), e.g., on the importance of a good formalism in biology for making predictions, framing hypotheses, revealing important features of the target system and making them salient,

[12] The following also holds for less traditionally "cognitive" capacities, like emotion regulation.

[13] A caveat: we do not want the most detailed or accurate model of that causal structure. We want a model that coarse-grains, idealizes, and is occasionally outright wrong, wherever those features are theoretically fruitful. We model a calculator with an addition-function, not an addition-except-where-the-calculator-errs function. So in the case of computational explanation, like causal modeling in general, accuracy with respect to causal structure is just one goal, tempered by others. I'll set this aside for now, but in Richmond (n.d.-c) I go into some detail about how representational thinking supports this idealization and coarse-graining.

and providing unity to the field. But it is already implicit in this that not just any language, even a highly descriptive one, will do, and (B) is required to ensure that our formalisms are functionally appropriate to our subject matter — along with our conceptual resources, they should facilitate *theoretically useful descriptions* of the causal structures we seek (Lazebnik, 2004). Cognitive science does not just need a language to *describe* the brain, but also to state its explananda and to frame relevant questions and hypotheses about those explananda. A formalism borrowed from particle physics might do a good job of describing the structure of the brain in many respects, but it would likely be difficult or impossible, within that formalism, to state explananda having to do with (say) an organism's capacity to memorize strings of words, or to frame hypotheses about the aspects of the brain's causal structure that allowed it to (say) reason deductively. A formalism that failed in these respects would need to be either abandoned or supplemented with conceptual resources that allowed it to do this work. These two desiderata, (A) and (B), will play the largest role in my discussion. (C), although it is an important part of any research program, is more nebulous. Among other things, (C) points out that the descriptions, predictions, explanations, and models that (A) and (B) allow us to give should ideally cognitively tractable for scientists, facilitate further investigation, have well-understood or clear properties, and the like. Together, what (A)–(C) would give us is a set of tools with which we can describe the causal structures in the brain that bring about its behavior and support its cognitive capacities (from A), understand those structures in relation to our explananda, and form relevant questions and hypotheses about them (from B), and work efficiently on those questions and hypotheses (from C). (In addition to these desiderata, it is also crucial that our formalisms and conceptual resources make it possible to *test* our theories. But this is essentially the problem of giving them empirical content, and I'll treat that in section 1.4.)[14]

My claim is that computational notions provide a set of formalisms and conceptual resources satisfying (A)–(C), and as such they contribute to the goals of cognitive science by providing

[14] I'm not claiming this is all that a formalism and its associated conceptual resources should do for a field of science. But (A)–(C) alone will serve to display much of the way computational explanations function and the reasons they succeed.

resources to explain how the internal structure of a system brings about its cognitive capacities. Since they do this well, they facilitate good explanations. And since those explanations hinge on fruitful and relevant descriptions of causal structures — not on the subsumption of a target system under the definition of *computer* or *computes* — then we need not worry about that definition and whether the brain satisfies it, any more than we worry about the definitions of *calculizer*, *virus*, *fluid*, or *race-car* in the earlier examples. The task, then, is to see how (A)–(C) are satisfied by computational notions and formalisms.

I'll start by focusing on formalisms. We need a formalism in which to capture the kinds of causal structures cognitive scientists seek — causal structures that explain cognitive capacities. There is no unique formalism appropriate to this task, and, for that matter, it is unclear how formalisms should be individuated. In particular, what counts as a computational formalism is not straightforward, and seems to depend on how tools from computer science are exported into new domains (Smith, 1999). The formalism of Turing machines is used in computational explanations, as are the formalisms of finite state automata and combinatorial state automata, the formalisms involved in describing perceptrons and artificial neural networks, on through more generic forms of description like wiring diagrams (e.g. Sejnowski et al., 1988), arithmetical operations (e.g. R. Devalois & Devalois, 1993), calculus equations (e.g. Shadmehr & Wise, 2005), and statistical functions (e.g., when a neuron is described as computing a Laplacian of Gaussian function, Egan, 1999, p. 192).[15] So I won't rigorously define "computational formalism." Instead I will lean on the way the notion of computation is used in cognitive science, and I'll allow that whatever, for cognitive scientists, counts as a computational description, is a computational description. The task is to see what those descriptions have in common that makes them suited to achieving the goals I've outlined.

What the formalisms above have mostly in common is that they invoke the devices built by computer engineers (e.g., wiring diagrams), the programs designed by computer programmers

[15] And to the extent that artificial intelligence informs cognitive science, its developments will introduce new and unpredictable computational formalisms (Kriegeskorte & Douglas, 2018). The concept of computation, as it figures into cognitive science, is 'open-textured' (Waismann, 1968) in at least this respect.

(e.g., neural networks), or the mathematical structures investigated in computer science (e.g., Turing machines and finite state automata). In using these formalisms, cognitive science describes the brain in terms borrowed from the science, engineering, or programming of computers, broadly construed. I'll condense this by saying it uses formalisms borrowed from the computing disciplines. These formalisms — call them the computational formalisms — are computational explanation's answer to (A). The conceptual resources that allow computational explanation to meet (B) and (C) are the ones attendant on the relevant formalisms, or that are otherwise drawn from the computing disciplines.[16]

The case to be made is that these formalisms and conceptual resources serve (A)–(C) well: that describing and conceptualizing the brain using them serves cognitive science's broader goals. So I will turn now to some of the ways that computational formalisms and their attendant conceptual resources serve those goals. One important feature of computational formalisms is their facility with functional abstraction. Functional abstraction highlights an aspect of a system component, usually described mathematically, that captures its contribution to the system's behavior at a higher level of abstraction. A paradigmatic example is naming high electron flow in a wire "1", and low electron flow "0" (Hillis, 1998, pp. 18–19). Any variation in the electron flow within either "1"-signals or "0"-signals disappears, along with the gradient between "1"- and "0"-signals, and all the wire's other features. In fact, the wire itself disappears. All that remains is the distinction we've selected as significant for our purposes — a distinction between 1 and 0. Because we've chosen a description of the wire under which it behaves predictably (we know the circumstances that will put the wire in a 1-state and the circumstances that will put it in a 0-state), we can exploit that distinction to build more complex functions like logic gates.[17]

There is no need to belabor the utility of functional abstraction for engineering, but it is important that it offers benefits in the reverse-engineering of the brain as well, particularly in a computational context. The saltatory action potential, e.g., lends itself well to a characterization in terms

[16] I'll return, in the next section, to borderline cases that may not be captured by my definition, like the arithmetical operations mentioned above.

[17] The examples here are simplistic for the sake of explanation. Functional abstraction is most commonly discussed in more demanding contexts, e.g. abstraction methods for managing complex databases.

of 1s and 0s; this was an explicit motivation for von Neumann's (Neumann, 1958) and McCulloch and Pitts' (McCulloch & Pitts, 1943) treatment of the brain in computational terms.

The story is now, of course, much more complicated. We don't treat neurons as logic gates but as (something at least as complex as) non-linear functions of weighted sums of inputs. But it is still a major goal of cognitive science to "decompose cognition into functional components," and to discover how the brain's activity at an "elementary" or neural level can be characterized so as to compose those functions (Kriegeskorte & Douglas, 2018). And to do this we need a way of abstracting from the complex causal profile of neurons (or ensembles of them, or brain areas) to well-understood mathematical functions. To be clear, functional abstraction is an unavoidable feature of the mathematical description of any physical system. The question is which mathematical formalisms to use. And what better formalisms than the ones for which we understand the implications most relevant to us? We know a great deal about the processes defined by computational formalisms: how fast they are, how many steps they take (if they are step-wise processes), how they scale to different inputs, how efficient they can be at what cost to accuracy, what they can do with and without recurrent steps, and so on (e.g. Kriegeskorte & Douglas, 2018, Box 3), and these are many of the same questions we have about the brain. So computational formalisms give us a toolkit for functional abstraction that is particularly well-suited to the questions we have about the brain.

Computational formalisms and the conceptual frameworks they bring with them also lend themselves to descriptions in terms of *algorithms* and *hierarchies*. Algorithms are functions strung together into (formally computable) sequences. They provide a clear and intuitive way of connecting a system's inputs to its outputs by describing the steps taken by the system in transforming inputs to outputs. That kind of description is precisely what cognitive scientists seek, as I suggested above: a description of the internal causal sequences that bring about cognitive capacities, the latter understood in terms of responses (or outputs) to environmental conditions and stimuli (or inputs). E.g., color-processing in early vision is modeled as an algorithm first summing responses from different types of cone, then weighting those sums, then adding and subtracting the weighted

values, and eventually plotting the results in a three-dimensional space (R. Devalois & Devalois, 1993; Mancuso et al., 2010). That is a description, in terms of an algorithm, of the way the brain turns a retinal input into a behavioral (or phenomenal) output.

Moving on from algorithms, hierarchies are processes that operate at more than one level of abstraction. One kind of hierarchical description is just an important kind of algorithmic description. Neural network models show us how a process can derive more and more abstract or high-level features of an input, e.g. an image, through a series of functions that finds its low-level features like lines and shapes, then intermediate-level features, and eventually high-level ones like object types (e.g. *dog* or *cat*). This was also an explicit goal of earlier, classical computational modeling (Marr, 2010). In computational neuroscience, this kind of hierarchical algorithmic description is essential to understanding how the brain makes categorizations, and especially how it proceeds from sensory stimulation to sophisticated high-level categorizations and behavior based on them. Computational frameworks drawn from neural networks (and, earlier on, other sources in the computing disciplines Marr, 2010) have helped illuminate the relevant brain processes by providing useful and increasingly accurate models of them (Richards et al., 2019).

The other kind of hierarchical description is *compositional*, rather than algorithmic. A compositional hierarchy is not a series of functions deriving higher- and higher-level features, but a hierarchy where a small set of simple functions compose more and more complex ones. E.g., the processes involved in different capacities may rely, at a lower level of their hierarchies, on a small "set of standard (canonical) neural computations: combined and repeated across brain regions and modalities to apply similar operations to different problems" (Carandini, 2012) (see also Carandini & Heeger, 2012). An understanding of this sort of hierarchy does a number of things for cognitive science. Understanding the simplest neural functions guides anatomical investigation into basic units and circuits and makes salient certain aspects of their causal structure (Carandini & Heeger, 2012). Without the simplicity and structure given by hierarchical thinking, it would be prohibitively difficult to connect cognitive neuroscience to basic physiology and anatomy, or generally to lower levels of brain organization. An understanding of how low-level brain structures

compose high-level ones also benefits modeling, since it reveals relevant and practical levels of description, particularly when the goal is (as is common) to model how high-level behavior results from low-level organization (Yamins & DiCarlo, 2016). For all these purposes, the benefits of computational formalisms are clear: their hierarchical properties are relatively well-understood; many computational formalisms are developed precisely for their ability to compose complex functions from less complex ones, especially a *small set* of less complex ones (particularly relevant when we consider canonical computations as above); and they are developed to create and make intelligible complex relationships at different levels of detail and abstraction.

The importance of hierarchical and algorithmic explanation, and the way computational formalisms accommodate them, is the last point I'll raise in support of computational formalisms being a good solution to (A). Moving on to (B), the assimilation of one system under the conceptual scheme developed for another system is a widespread and natural part of science (Dunbar, 2002; Nersessian, 2002), and conceptual schemes from the computing disciplines have been particularly useful ones in which to assimilate the brain. In fact, the discussion so far has already shown that computational formalisms and the conceptual frameworks attendant on them are well-suited to frame explananda to do with cognitive capacities, and to pose questions and hypotheses about the processes that bring them about. E.g., I mentioned the that neural networks provide frameworks for thinking about how the brain's causal structure supports the derivation of object categories from lower-level features of a stimulus. But it is worth noting a few more cases. E.g., considerations of algorithmic complexity — an important concept in computer science — drive discussions about the appropriateness of Bayesian models of the brain (Kwisthout & van Rooij, 2020). Considerations of computational efficiency — important in computer science and computer engineering — drove early debates in cognitive neuroscience (McClelland et al., 1986), and considerations of "computational cost" drive current discussions of navigation and route planning (Daniel et al., 2015). It is because we think of the brain in computational terms that we investigate its canonical operations, as above. It is because we understand the properties of recurrent connections in neural networks that we look for recurrent connections among neurons (Richards et al., 2019). The search

for the brain's learning rules and functional architecture is spurred and supported by thinking of it in terms borrowed from computer science and neural network engineering (Richards et al., 2019). So assimilation into the conceptual schemes of the computing disciplines provides many concrete benefits for generating hypotheses, and for understanding our causal models of the brain in relation to their explananda. If this seems to belabor the obvious, recall that the take-away is not that it is useful to think of the brain as a computer — that *is* obvious. The take-away is that we can make sense of this fact without claiming that the brain *is* a computer, and without entering into vexed questions about what exactly that means. All we need is to understand how computational formalisms are used to describe causal structure, and why they are so useful for this task. That's what I've described above, with no need for metaphysical commitments; in their place I have appealed only to complex but scientifically common-place considerations about explanatory goals and the tools with which we pursue them.

To finish with a familiar point about (C), computational explanation makes it possible, and relatively easy, to *build* models. Compare Kriegeskorte and Douglas (2018): "only synthesis in a computer simulation can reveal what the interaction of the proposed component mechanisms [of some theory] actually entails and whether it can account for the cognitive function in question." It is a common refrain in the history of cognitive science that computational models make hypotheses clear and specific, and they do this partly by making them buildable using our current technology (e.g. Churchland & Grush, 1999; Pylyshyn, 1984; Samuels, 2019; Sejnowski et al., 1988). To grasp the significance of this one need only imagine a theory of the visual cortex as a convolutional neural network, but imagine it proposed 50 years ago. The benefits this theory has because of current computing technology (to do with prediction, ease of understanding, the availability of proofs of concept, our familiarity with the model and an intuitive understanding of what it says about its target system, etc.) are the benefits I'm claiming computational models have in general because of their buildability, and because of the familiarity we therefore have with them.

That's all I'll say about (A)–(C). To summarize, computational explanations, by drawing formalisms and conceptual resources from the computing disciplines, support cognitive science by

providing: explanatorily relevant functional abstractions of the brain's causal structure; descriptions of that causal structure in the fruitful terms of algorithms and hierarchies; tight connections between our descriptions of the brain's causal structure and cognitive science's *questions* about that causal structure; relevant and fruitful ways of framing answers or hypotheses concerning those questions; and, more generally, explanations of how the brain supports cognition that are natural, powerful, and sensitive to our specific interests in cognition and our existing knowledge of the brain. And it does all of this with relative efficiency by drawing on well-understood/understandable, well-established, and deeply-ingrained conceptual frameworks. There is more to be said about the details here, but it is a benefit of the discussion so far that it relies only on uncontroversial features of computational formalisms and concepts. It is not that *on a certain tendentious way of thinking about computation* we can do without a metaphysics of computation. To begin to answer our questions about computational explanation, and to do so without building a metaphysics of computation, we need only appeal to the features of computation that we are all aware of.

The goal was to sketch a view of how computational explanation works and why it is so successful in cognitive science. The view, more succinctly, is this: computational explanation works by using computational formalisms and the conceptual resources attendant on them to construct process models that capture the causal structures in the brain that bring about its cognitive capacities. If this is what computational explanation does, it requires no assumption that the brain is a computer, much less a theory about what it is to be a computer. This account reveals, at a general level, how computational explanation works, and also why it is so successful: it serves the purposes of cognitive science exceptionally well. The account also shows us what makes computational explanation *distinctive* as a mode of explanation — not the subsumption of the brain under a special definition or category, but the powerful suite of tools, resources, and concepts it draws on to serve the particular purposes of cognitive science.

1.4 Objections

The bulk of the Pragmatic Approach is on the table. In this section I'll consider some objections, focusing on ones that will let me sharpen the approach a little further. To bring out the most pressing worry, let's start by returning to a version of the triviality problem.

1.4.1 Empirical content: another triviality problem?

The triviality problem can't arise in its original guise: it attacked the definition of 'computer,' or the criteria for membership in the category COMPUTER, which are not involved in my view. But it might be resuscitated along the following lines: if we don't know what computational explanations say about their target systems, we haven't fully understood how or why computational explanation works. I've described computational explanation as a certain kind of modeling practice, but I haven't said how to tell what a given computational explanation concretely says about its target system. That is, I haven't yet placed any constraints on the *empirical content* of computational explanations. And if there are no constraints on the empirical content of a model, then why can't we interpret it as saying whatever we like? Why can't I give a computational process model of a rock as performing addition, and say that the model is correct as long as the rock runs through time-slices that correspond to the stages of the model's addition algorithm? This was a long way to come, just to end up back at the original problem.

But the problem is only apparent. Computational explanations say something about the causal structure of their target systems, and two constraints ensure that what they say about this structure is non-arbitrary. First, this version of the triviality problem is a special case of a more general problem: in virtue of what does any model say what it says about the system it says it about? I don't propose to answer this question here, but on the view I've defended, the question of the content of a computational explanation is just an instance of this more general question of model reference (Frigg & Hartmann, 2018; Frigg & Nguyen, 2018). And we can be sure that it is not arbitrary what scientific models in general (and therefore computational explanations in particular) say about their

target systems. Whatever your account of scientific models, you have to explain their empirical content somehow, and there is no reason that computational explanation in particular would be excluded by whatever account you adopt. Note also that the problem of model reference is not generally solved by criteria for a target system's membership in a particular category. The revision of the triviality problem I'm considering would apply equally well to models of the solar system constructed in calculus equations, and the problem of why those models say what they say about their target systems is not solved by criteria for being a *calculizer*. Likewise for models of traffic from fluid dynamics, models of disinformation from virology, and so on. So the problem of model reference is no argument against the Pragmatic Approach. If you think models can have non-trivial empirical content, there is no special reason to worry about computational explanations.[18]

The first constraint, then, is the general theory of model reference — whatever we say about *that* will apply to and constrain the content of computational models. The second constraint comes from existing scientific knowledge. The goal of a computational explanation is to describe the causal structure that brings about a system's capacities. Not just any empirical content is appropriate for this task. If a neuroscientist held that her model of memory was confirmed by connectome data because that data revealed that the brain had *just some mapping* to her model, she would be laughed out of the lab meeting. But many more specific mappings would also be dismissed. What counts as an appropriate mapping of model to brain (appropriate empirical content) depends on background neuroscientific knowledge about (e.g.) which aspects of brain activity are involved in the tasks she's modeling, which components of the brain are causally efficacious in the right ways, and so on. To explain a memory task, she might propose that synaptic weights correspond to certain terms in her computational model. This would be appropriate if synaptic weights were causally implicated in memory in the required way, but it would be inappropriate if Gallistel and King (2009) were right that synaptic weights cannot bear significant responsibility for memory.

[18]Others have made similar points. Matthews and Dresner (2017) argue that triviality arguments about computation have the same structure as triviality arguments about any attribution of numerical properties to physical systems, and so cannot hold. The advantage of the Pragmatic Approach is that it says something positive about computational explanations and their success, and shows what exactly is wrong with the triviality problem: it challenges computational explanation only under the Metaphysical Approach, which is — I've suggested — mistaken.

Empirical constraints on model interpretation are familiar to cognitive science — e.g., see discussions of "mappable" models (Yamins & DiCarlo, 2016) or "explanatory mechanisms" in model building (Blohm et al., 2020). For more concrete examples, consider debates over whether neural spike *rates* or *timings* are causally efficacious in the brain, and which should be the target of our models (Brette, 2015). Or see debates about modeling population activity vs modeling individual neurons and their connections (Barack & Krakauer, 2021). The empirical content of computational models is partly constrained by the diverse empirical considerations that constrain the empirical content of all models, as we should expect.

1.4.2 (Yet) another triviality problem

The recipe for a triviality problem is to take something for which we need criteria, and to show that some proposed criteria encompass too many things. On the Pragmatic Approach there is nothing it is to be a computer — we have and need no criteria for this. But there seem to be circumstances where computational explanations are appropriate, and ones where they aren't. There appear to be criteria for the *appropriateness* of computational explanation.

So, why is it not always acceptable to use computational explanations, regardless of one's target system or explananda? In one sense, it is! We should have no qualms with someone to whom computational formalisms and conceptual resources derived from the computing disciplines are helpful for explaining (say) planetary systems or the weather, because accommodating this does not require a revisionary metaphysics according to which planetary systems and weather patterns are computers. We should note, however, that computational explanations are in fact not helpful in most cases, at least to us in our current context. They do not help us make sense of the behavior of the weather, nor rocks or walls or pails of water. The use of computational explanation should not be barred anywhere a priori, but in practice it is not appropriate in every case or context. As I've described computational explanation, it introduces a particular set of resources that are useful for a particular set of goals, in a particular scientific context, given particular constraints imposed by our target systems and the nature of our inquiry. There is no reason to expect this set of tools to be

useful for *just any* purpose, in just any context, with just any constraints. We should perhaps expect computational explanation to be particularly successful for systems that have undergone a design process (including design by selection) to create a structure that efficiently generates appropriate outputs from inputs, because it is from disciplines creating and studying that kind of system that computational explanation draws most of its resources. But I won't pursue this here. The point is that although there is no reason to bar it a priori, a rock is unlikely to receive a successful or fruitful computational explanation.

1.4.3 One last triviality problem

What about the category *computational explanation* itself? What makes something a computational explanation? To frame this explicitly as a triviality problem: why doesn't every explanation count as a computational explanation? An answer is implicit above: where formalisms and conceptual resources drawn from the computing disciplines are used to meet explanatory needs like (A)–(C), you have a computational explanation. Otherwise you don't. Consider a model of a two-body system given in calculus equations. (A) does not appear to be met — though it presumably could be in *some* model of the system. But even if it were, (B) is met only to a lesser degree, if at all — the conceptual resources required to understand and form hypotheses about the system do not come from a computing discipline. And (C) is not met at all: the heuristics and background knowledge required to understand and efficiently inquire about the model and system do not come from a computing discipline. So even if a two-body system were given a process model in a paradigmatic computational formalism, it would not be given a computational explanation.

We might worry about borderline cases, e.g., computer models of evolutionary processes that assume a sort of optimality to natural selection, and look for algorithms to achieve it. Say we have such a model, for which (B) and (C) are met to a large degree by conceptual resources from computing disciplines, but not to the same degree as in a typical computational model of (say) visual processing. In that case we should be happy to say the explanation is closer to a computational explanation, or more of a computational explanation, or is a more paradigmatic computational ex-

planation. There is no reason to expect computational explanation to be a binary category. Since it is defined by the use of tools from the computing disciplines, those tools and those disciplines being fuzzily defined themselves, we should expect a fuzzy spectrum rather than strict criteria for counting as a computational explanation. This does not make it any harder to understand computational explanations or the source of their explanatory significance. That source was not their belonging to some strictly-defined category, it was their use of certain resources to meet particular needs. Some of those resources can be present — and therefore relevant to the explanations' force — while others are not, or are only to limited degrees.

I can now address an issue I postponed earlier. Some of the computational explanations I've mentioned, e.g. the ones to do with color vision, don't use formalisms drawn from computing disciplines. They use arithmetic. A certain set of retinal ganglion cells are described as computing S – (L + M), where the letters refer to the responses of different cone types. These explanations do, however, conceptualize the retina as following algorithms, one of the computational conceptual resources I pointed out above. And they may draw on knowledge from the computing disciplines, e.g., to do with the efficiency of different algorithms, the use of population-coding, data compression, and information-processing more generally (e.g., see Jameson et al., 2020, passim). If this terminology is just window-dressing on an explanation that doesn't make significant use of formal or conceptual resources from the computing disciplines, we're looking a non-computational explanation. But to the extent that this terminology and these conceptual resources contribute to the explanation of color vision, to the extent that thinking in these terms serves our explanatory purposes, we have a case of computational explanation. There are more and less computational ways of describing color vision, and, again, some will only weakly count as computational explanation. But there is no reason to demand that they count as full-blooded computational explanations when we have a good explanation of their function and success that doesn't require it.

I should pause here to note a caveat. I did not set out to explain *what computational explanation is*. I set out to explain *how and why computational explanation works*. To see how something works, you usually don't need to know what it is to be that thing. If you doubt this, try asking

your mechanic for a rigorous definition of "engine," or a metaphysics of that category. So the above is intended to clarify the scope of my account and how it handles less paradigmatic cases — not to sharply define the category COMPUTATIONAL EXPLANATION. By way of illustration, consider Cisek's (Cisek, 1999) argument that cognition is non-computational because it consists largely of control processes that are not well-captured by classical computational thinking. On the Pragmatic Approach this does not raise a question of whether cognition is *really* computational or not, nor, to the present point, the a question of *whether the resources of control theory are really a part of computational explanation or not*. Instead, it raises the question of whether certain useful resources are neglected in cognitive science, what work they could be put to, and how to introduce them to do that work. There is just no need to define computational explanation so as to include *or* exclude explanations drawing on control theory. There is, instead, a need to investigate control theory and its potential usefulness in cognitive science, and to introduce it where it will be helpful. That project is ongoing (Richards & Lillicrap, 2022), but has nothing to do with the metaphysics of computation or the definition of "computational explanation."

1.4.4 Metaphysical appendices, naturalism, and objectivity

One final concern before I conclude. Consider a possible rejoinder to the Pragmatic Approach. You might try holding on to a metaphysics of computation while retaining the benefits of the Pragmatic Approach by giving the above account, and simply adding an appendix that says: whatever systems receive legitimate computational explanations according to the Pragmatic Approach *are thereby computers*. This lets us classify the brain as a computer without requiring our metaphysics to do any heavy lifting. Note that this approach — the Metaphysical Appendix Approach — accepts that the metaphysics of computation are irrelevant to understanding computational explanation. The proponent of this view just has some other reason to want a metaphysics of computation. (Maybe it's the kind of concept that *just can't fail* to correspond to a property, even if the use of the concept relies not at all on that property.) So they have accepted my main conclusion: if you want to understand how and why computational explanation works, you have no need for a metaphysics

of computation. But then it's hard to think of what would motivate the Metaphysical Appendix Approach. It's hard to think of a context in which the metaphysical appendix will play an important role, except perhaps if we're listing and describing all the properties our world contains — but recall that this list is, by hypothesis, irrelevant to science and the philosophy of science. Not that I have any problem with this variety of stamp collecting, but it has no role in answering the kind of question I've taken up here.

And note: harmless though they may seem, appendices are liable to burst, and an approach that insists on a metaphysical appendix leaves itself open to complications. The resulting metaphysics of computation would be "stancey" and observer-relative. It is likely to be graded and fuzzy as well, given the discussion in section 1.4.3. These are not good objections, because on the Metaphysical Appendix Approach there is no reason to require a metaphysics of computation to be objective, observer-independent, etc. Perhaps a metaphysics *that is relied upon in science* should be all of those things, but one that exists only as an appendix for some other purpose cannot be held to these standards. One might as well criticize the category THE LEAST TASTY BUBBLE GUM for being "stancey" and subjective. But these objections, misguided as they are, bring with them a dialectical context — the context of the question, "what is it to be a computer," where they can be confused with good objections, or even grounds for rejection, and lead to the kind of debate we see between, e.g., Brette (2022) and Richards and Lillicrap (2022).

I've taken the appendectomy because these complications are likely. Consider the widely-accepted desiderata that a metaphysics of computation be *naturalistic* (Sprevak, 2019, p. 177) and *objective* (Piccinini, 2015, p. 12), in that what counts as a computer should not rely on human beliefs or purposes. If you take the Metaphysical Appendix Approach, you will meet neither desideratum: what counts as a computer for you will depend heavily on human beliefs and purposes. But instead, you could leave aside the metaphysics altogether. It is not even a prima facie objection to an account of *how computational explanation works* to say that its working depends heavily on human knowledge, explanatory goals, and so on. In fact, we should expect this of every form of explanation, since explanations serve various human purposes and to do so must interact in

particular ways with our interests and background knowledge. Computational explanation depends on human goals and purposes in just the same way that, e.g., fluid dynamical models do, whether they are applied to fluids or traffic. If there is a related problem for the Pragmatic Approach it is just to show that computational explanation is nonetheless constrained so that its explanatory significance is not undermined. And, as I've argued above, computational explanation on the Pragmatic Approach *is* so constrained — just not by metaphysical considerations.

1.5 Conclusion

Let me summarize. Computational explanation in cognitive science works by using formalisms drawn from the computing disciplines, and the conceptual resources attendant on them, to construct process models that capture the causal structures in the brain that bring about cognitive capacities. It is successful as a general strategy because it serves the needs of cognitive science, and it is successful in specific instances, like any modeling practice or strategy, because (or if) it accurately describes the causal structures that bring about the behavior or capacity under investigation, and in a way that meets the various standards we have for scientific models and explanations. On the view I've described, there are limits on which computational explanations appropriately apply to which systems, preserving the explanatory significance of computation. Questions remain about the kinds of formalism computational thinking introduces, or should introduce, into cognitive science. But these questions should be answered along the lines of the Pragmatic Approach: not through the metaphysics of computation, but by a careful look at the resources that the explanations of interest involve and the goals they are intended to serve.

So far I've merely built up the Pragmatic Approach, letting the sense it makes of computational explanation stand as evidence for it. And I'm content to have simply gotten this approach on the table, ready for further discussion and refinement. It makes a sharp contrast to conventional approaches (e.g. Brette, 2022; Piccinini, 2015; Richards & Lillicrap, 2022; Shagrir, 2022) — a contrast that illuminates the assumptions of those approaches and (I've argued) their mistaken focus. But in the introduction I promised you a further argument against the Metaphysical Approach.

That argument is this: if there is a working account that does without the metaphysics of computation, then to justify an account that accepts such a metaphysics, *you would need a reason to think that metaphysics is necessary*. The necessity of this metaphysics is rarely argued for, but the Metaphysical Approach posits and focuses on an entity — a kind, property, or category: COMPUTATION. An argument for this approach must specify some desiderata that the Pragmatic Approach doesn't achieve, but that the Metaphysical Approach does. Or it must give some other reason we should expect this property to exist *and* to be relevant to cognitive science. Otherwise the Metaphysical Approach posits and spends time investigating the property for no apparent gain — it is explanatorily redundant and contributes nothing to our understanding of scientific explanation.

There is another way of coming at this point. To take the Metaphysical Approach is not only to posit a certain property, *computation*; it is to understand cognitive science and scientists as committed to the existence of that property, and to a specific understanding of that property. But these are commitments that cognitive scientists don't intend to, or appear to, make themselves.[19] Take just two examples of how mainstream cognitive scientists see their practice. Yamins and DiCarlo (2016) understand their preferred type of computational explanation as a way of "formalizing knowledge about the brain's anatomical and functional connectivity" so as to explain its cognitive capacities — not a metaphysical claim at all, and quite in line with the view I've defended here. And Richards et al. (2019) take deep learning and the computational explanations associated with it to involve the application of an explanatory/investigative "framework" involving specific types of models and hypotheses, principles about the causal structure of the brain, and strategies that the history of neural networks suggests for understanding complex systems. This is well-captured by the view I've defended, but to hold on to the Metaphysical Approach we would have to impose

[19] And cognitive scientists who do seem to make that commitment are quick to fall back to a pragmatic approach when issues like the triviality problem arise. A complication is that talk of *what computation is* may provide a useful framework for thinking about one's approach to computational models (thanks to Richard Lange and Rosa Cao for making this point in conversation). Thinking about what computation is could be just a way of thinking about computational models. I could hardly object to this if it really was just a methodological trick, rather than a substantive metaphysical commitment. But I take it that this framing almost always does involve metaphysical commitments. Regardless, if you can avoid those commitments, and see questions about the metaphysics of computation as a metaphysically non-committal shortcut to questions about computational explanation, we should have nothing to disagree about.

on this area of cognitive science commitments that it does not appear to make and that offer no advantages over an approach that sticks more closely to scientific practice itself.

This is just one count on which the Pragmatic Approach is preferable to the Metaphysical Approach, but it reflects a broader range of considerations of significance to philosophers of cognitive science. To build the bridges between cognitive science and philosophy that most philosophers desire, it will be important to avoid, as far as possible, foisting the assumptions and definitions of one field onto the other. The Pragmatic Approach avoids at least one such foisting that the Metaphysical Approach does not. And, in fact, my version of the Pragmatic Approach does so by focusing on the context, practice, and function of computational explanation for cognitive scientists — another necessity for the bridge-building that philosophers have their hearts set on.

To conclude, computational explanations provide a powerful, and crucial, lens on the brain. Philosophers of cognitive science, and many cognitive scientists themselves, have been duly impressed by the computational lens, but have failed to see it as a lens, instead understanding computation as a property of the brain itself. This is a natural enough mistake — a good lens is not perceived; it is perceived through. But if we forget we're looking through a lens, we will vastly misunderstand the things we see through it. For that matter, thinkers of a certain sort are liable to leave the lens on, turn to a chunk of rock, and shudder to discover that it has all the brain's computational properties too.[20] When, instead, we understand computation as a lens, we begin to see what it does to its target, what it occludes and makes salient, what it adds, what it blurs, what it brings into focus, and how, in turn, it makes the brain intelligible as the organ responsible for the mind.

[20] As in Brette (2022), Chalmers (1996), Milkowski (2013), and Piccinini (2015), for just a few examples.

Chapter 2: Computational Externalism

2.1 Introduction

The goal of this paper is to establish a thesis of long-standing significance for both cognitive science and the philosophy of mind. That thesis is *computational externalism*: the brain does not have a computational structure in itself, but only in conjunction with things outside of it. In section 2.2 I will outline computational externalism, with special attention to its implications for cognitive science. In section 2.3 I will argue for computational externalism through a case study in the evolutionary biology and neuroscience of color vision, which raises a challenge internalism likely can't meet. And in section 2.4 I'll discuss the implications of computational externalism for cognitive science itself, specifically regarding the explanatory relevance of connectomics, the importance of naturalistic stimuli, and the question of whether "neuroscience needs behavior" (Krakauer et al., 2017).

2.2 What is computational externalism?

The brain does not have a computational structure in itself, but only in conjunction with things outside of it. What exactly does that mean? Let's start with *computational structure*.

2.2.1 Computational structure and computational explanation

I will use a notion of computation that is maximally sensitive to its function in cognitive science, and minimally committal otherwise. My only assumption will be that the notion of computation must support the kind of judgments, inferences, and classifications that cognitive scientists make as they give and justify computational explanations, and I will use the barest notion of computation that serves that purpose. So, e.g., I won't assume that computation must involve

representations, teleological functions, or the like.

What does this minimally-committal notion of computation involve? Most importantly, it means that there is a difference between a system being *modeled* computationally and *performing* or *implementing* computations — or, as I prefer to put it, between a system being modeled computationally and being *genuinely explained* computationally.[1] We model all kinds of systems computationally, e.g. the weather, but cognitive scientists are doing something very different when they give computational explanations of the brain. Again, though I put this differently elsewhere, we can say that the brain *implements* computations in a way that the solar system doesn't. There are many theories of implementation — of what a system must be or do to implement a computation. At the very least, it seems necessary that the system go through a process that approximately mirrors the stages of the proposed computation (Cummins, 1991; Egan, 2010, 2014; Pylyshyn, 1993). If I say a system is adding pairs of numbers, the process it undergoes had better have parts (states or components) that can be identified as the addend_1-part, the addend_2-part, and the sum-part, and it must pretty reliably transition from addend-states to sum-states, or use the states of the addend-components to determine the state of the sum-component. Otherwise, even if the system's behavior can be predicted by the addition computation, the computation isn't *going on inside the system* — we have a case like the weather, not the brain. This minimal condition is all that I will assume cognitive scientific practice commits us to. There are many (and contentious) views of what else computational explanation involves, but rather than enter into that debate I want to show that *this minimal notion of computation is all that's required to establish externalism*. So I won't assume that computing systems have any features other than this: they have parts or components that approximately map to the algorithms the system implements.[2]

By contrast, note how much easier it is to argue for externalism if you assume that compu-

[1] See Milkowski (2013, Chapter 1), Piccinini (2015, Chapter 1), Richmond (n.d.-b) for three ways of cashing this out.

[2] I'm not concerned here about whether this generates a triviality problem, because I think that problem can easily be solved on a more sophisticated understanding of computational explanation (Richmond, n.d.-b), and because I'm not arguing that this minimal notion of computation is *all computation is*. I'm claiming this minimal notion is enough to generate an argument for externalism. I'll also note that, if you are skeptical of even the mapping constraint as general constraint on computation, I think you will at least agree that it is operative in the case study I consider in section 2.3, which is all I will ultimately need.

tations are defined partly by the representations they operate on. Then, if a state's identity as a representation changes depending on its environment, in different environments the computations that state figures into will also be different (Fletcher, 2018; Peacocke, 1994, 1999; Piccinini, 2008, 2015; Piccinini & Shagrir, 2014; Rescorla, 2013; Shagrir, 2001, 2018; Shea, 2013; Sprevak, 2010). An argument for externalism independent of these representationalist assumptions will be more convincing (since it will depend on fewer controversial assumptions), and will reveal something unique about the foundations of computational explanation.

My minimalism about computation also makes my version of externalism somewhat more radical than the view I just glossed. That view doesn't understand the *syntactic* structure of a system, or the bare *causal* structure captured by computational explanations, to depend on the environment. Rather, two systems compute something different because their representations represent different things — a difference that percolates up to the identity of the computations those representations are involved in. The causal structure of the systems, insofar as computational explanation aims to capture it, is unchanged. But the minimal conception I'm using means that my argument for externalism must be an argument for syntactic externalism, or, as I'll call it, *causal structure externalism*. Since I assume only that computational explanations describe a certain kind of causal structure, my externalism means that *the brain's causal structure, insofar as it is captured by computational explanations, is not something the brain has in itself, but only in conjunction with things outside of it*.[3] I'll say more about what this means, and how it could be true, in a few pages. But of the few theorists who accept causal structure externalism about computation — as far as I know, just Bontly (1998), Horowitz (2007), Shagrir (2001, 2018), and Shea (2013) — all but Bontly argue for it on the basis of a representational account of computation, and Bontly invokes teleological considerations instead. These authors take the argument from the last paragraph one step further, arguing that what a system counts as representing can actually affect the causal structure of the system insofar as computational explanations are supposed to capture it. But here as well it would

[3]This version of externalism also entails that a formal or mapping account of computation does not suffice to establish internalism, as is sometimes suggested (see especially Egan, 1991, 1995, 1999, 2010, 2014). And nor does a narrow account of content establish internalism (as in Butler, 1998), since externalism does not depend on any facts about content in the first place.

be more convincing if we had a less committal basis for causal structure externalism — one that doesn't hinge on commitments to or accounts of representation or teleological function.

2.2.2 Internalism and externalism

The brain does not have a computational structure in itself, but only in conjunction with things outside of it. We know what computational structure is: for my purposes, it's the causal structure that computational explanations aim to capture. But so far I've glossed over this business of "in conjunction with things external to it." Here are a couple ways we could ask the question of externalism:

> Does a system's computational structure *depend on* features of the world external to it, or do features of the system itself *fully determine* its computational structure?
>
> Is it *in virtue of* its internal properties alone that a system computes what it does, or does it compute what it does partly *because of* the world external to it?

These are questions of classification, not causation. The "determine"s and "because"s and "in virtue of"s and "depend on"s refer not to the events that *brought it about* that a system computes what it does, but to the features that *categorize* the system as computing what it does, the features that make attributions of those computations correct. If those features include ones external to the system, externalism is true. If they do not, internalism is true.

But there are different ways to make precise the "determine"s and "because"s and so on. See, e.g., the two definitions described by Egan:

> [Internalism] in psychology is the claim that psychological states **[R:]** are taxonomized without *essential reference* to the environment of the subject possessing them; in other words, they **[S:]** *supervene* on the subject's intrinsic, physically specifiable, states. (Egan, 1994, p.258, emphasis mine)

The two definitions, S and R, are generally taken to be equivalent. When a specific formulation is required, S is usually preferred (Chomsky, 2000; Egan, 1992, 2003, 2010; Fodor, 1981; Sterelny,

1990; Stich, 2010). I think this is a mistake. Not only are the definitions distinct; R is the more useful one if we're interested in connecting the philosophical debate over externalism to cognitive scientific practice.

This could be derived from the case study to come (R will show its usefulness and S its uselessness in that case), but let me clear this up now rather than pausing in the middle of the action for a terminological discussion. Consider two rough taxonomic principles:

T1 A system's computational structure is partly determined by features of its current environment.

T2 A system's computational structure is partly determined by features of its current environment, *and* by features of all other environments that are of interest to cognitive science.

Both make essential reference to the environment, so they are externalist by the standards of R. But T2 also ensures supervenience on internal properties: every environment that cognitive science might be interested in moving the system to, or understanding its computational structure in, is already taken into account when the system's computational identity in its *current* environment is determined. As long as T2 doesn't privilege the current environment in any way (and lets stipulate that it doesn't), the computational structure of a system will not change as the system is moved from environment to environment.[4] So the principle is internalist by the standards of S, despite being externalist by the standards of R.

If T2 seems odd, compare the widely held desideratum that a theory of cognition should predict an organism's response to "arbitrary stimuli" (Rust & Movshon, 2005; Yamins & DiCarlo, 2016). Conventional cognitive science wants to understand the brain's response to a wide range of environments. But if it did this by allowing the brain's computational structure to change between environments, as in T1, we would have a problem. Organisms would have a different computational structure/computational explanation in different environments, and in different environments we

[4]That is, supposing we ignore environments that are, by hypothesis, of no interest to cognitive science. To reintroducing these environments as part of what determines a system's computational structure would be to flout my original commitment to understanding how computation works *in cognitive science*, and to working with a notion of computation that is maximally sensitive to cognitive scientific practice.

would need a different model to capture that different structure and explain the organism's behavior. The problem is that it is widely accepted in cognitive science that, even if many diverse models are our starting-point, "ultimately, one hopes *to integrate all these models into a single theory* that can predict neuronal and population responses to any arbitrary stimulus" (Rust & Movshon, 2005, p. 1647).[5] So if the environment helps to determine a system's computational structure, *it shouldn't do so in a way that requires a new computational structure for a system as it moves between environments*. That means that if environments other than the organism's current one play any role in determining its computational structure when it is *in* those environments, they should do so by playing a role in determining its *current* computational structure — or else we have the proliferation of unintegrated models that cognitive science wants to avoid. So T2 relates much more closely to cognitive scientific practice than T1.

This isn't quite an argument for R: so far, the only difference between S and R is semantic. One calls T2 internalist and the other calls it externalist. The problem is that if T2 or something like it is operative, S lets us get out of the externalism debate — the debate over whether the environment helps to determine the brain's causal structure — *without having to decide whether the environment helps to determine the brain's causal structure*. Either it doesn't, in which case internalism is, of course, true. Or it does, in which case — as long as T2 or something like it is operative — internalism is true again, since under T2 environment-shifting won't affect an organism's computational structure. R, on the other hand, counts a taxonomic principle as internalist or externalist on grounds that carve the debate more cleanly. On R, a principle is externalist just as long as it gives the environment *some role* in determining a system's computational structure, some place in the considerations that categorize it as having its particular computational structure. So to establish externalism we have to engage with the central question that S could bypass — whether the environment helps to determine the brain's causal structure.

R is also more relevant to cognitive scientific practice. The place where externalism is most

[5]See S. V. Shepherd and Platt (2010, p. 526) for an expression of this sentiment in the case of attentional mechanisms in primates, and the task of explaining their operation across natural social settings and artificial laboratory ones.

obviously relevant to cognitive science is the perennial debate about the relevance of ethology to neuroscience, recently revitalized by Krakauer et al. (2017), and followed up by a number of further papers, e.g., Niv (2020) and Cushman (2020). The question is whether we can understand a cognitive system if we neglect to carefully consider its environment and its behavior in that environment. For cognitive scientists this is a methodological issue rather than a substantive one. Should we worry about designing ecologically valid experiments? Must we do detailed behavioral analysis to understand brain data? Can we learn anything about the computational sources of behavior if we don't start from that behavior's environmental context?[6] But the substantive question bears on the methodological one: if externalism is true, our experiments and theories about the brain's computational structure had better take into account the relevant features of its environment.

But if there is a substantive question underneath these methodological questions, one that promises to shed light on them, S would tell scientists to answer it by running around a bunch of different environments and seeing whether their target system's computational structure changes. The problem is that *they wouldn't know whether it changed unless they had a way of ascertaining what that computational structure was in the first place*, i.e., unless they had principles or some basis for ascertaining a system's computational structure. And once they had a way of ascertaining its computational structure, it's unclear what new information any implications about environment-shifting would add. So the externalism that makes an immediate difference to cognitive science is one concerned with the *basis* of our computational ascriptions, or the reasons it is categorized as having the computational structure it does. Far less important are the consequences for environment-shifting systems, because not only does the latter depend on the former, but the latter also seems to be immaterial once we have the former. Because of this, and because of the previous point about how S and R carve up the debate, I will be working with externalism as defined by R.[7]

[6] See the papers in Platt and Ghazanfar (2010) for examples of work explicitly premised on affirmative answers to all these questions.

[7] Note that I haven't argued against S *altogether*. I have wanted only to show that for my purposes and given my questions, R will be more relevant and illuminating.

2.2.3 How could computational externalism be true?

Apart from the four causal structure externalists I mentioned above, there is a broad consensus against causal structure externalism — a consensus that a system's causal structure, insofar as computational explanations aim to capture it, could only depend on features of the system itself, particularly the structure and relations of its physical parts (Shea, 2013). To be sure, it is hard to see how the property of *having a certain internal causal structure* could be determined by something's environment. (Of course the environment might *cause* something to have its internal causal structure, but, as I've just discussed, that isn't the sense of determination involved in the externalism debate.) So I want to spend some time illustrating causal structure externalism with a toy system, so that you know what to expect from my argument. I won't argue in detail for externalism about the toy system. The point is to see *what it would mean* for externalism to be true, and to provide a schema for understanding the more complex case study to come.

The first thing to note is that not every bit of causal structure counts as computational. When we ascribe a computational structure to a system, we *coarse grain* its causal structure. The question, given R, is whether that coarse-graining is done solely on the basis of the system's internal structure, or whether the environment plays some role in determining the correct coarse-graining. Consider a simple network like the one in Figure 2.1, with three input nodes, A, B, and C. A is excited by light around 400nm, B is excited by *ultraviolet light* around 900nm, and C is excited by light around 600nm. They pass their signals on, with connection strengths denoted by the w's, to an output node. In the diagram, green arrows indicate excitatory connections, red arrows inhibitory ones. The output node incorporates these signals, implementing a simple threshold function over them: if its input is greater than 0 it fires; if not it doesn't. There are two complications to this structure. First, in the absence of ultraviolet light, *B is excited by C*. It sums its input from C and passes on that same value as output. But B contains a mechanism that *deactivates* the synapse with C as soon as B is photoelectrically excited. So in the absence of ultraviolet light, B will respond just as C does for any input. But in the presence of ultraviolet light, C won't affect B at all, and the activity of each will be determined by the incoming wavelength spectra. The other complication is

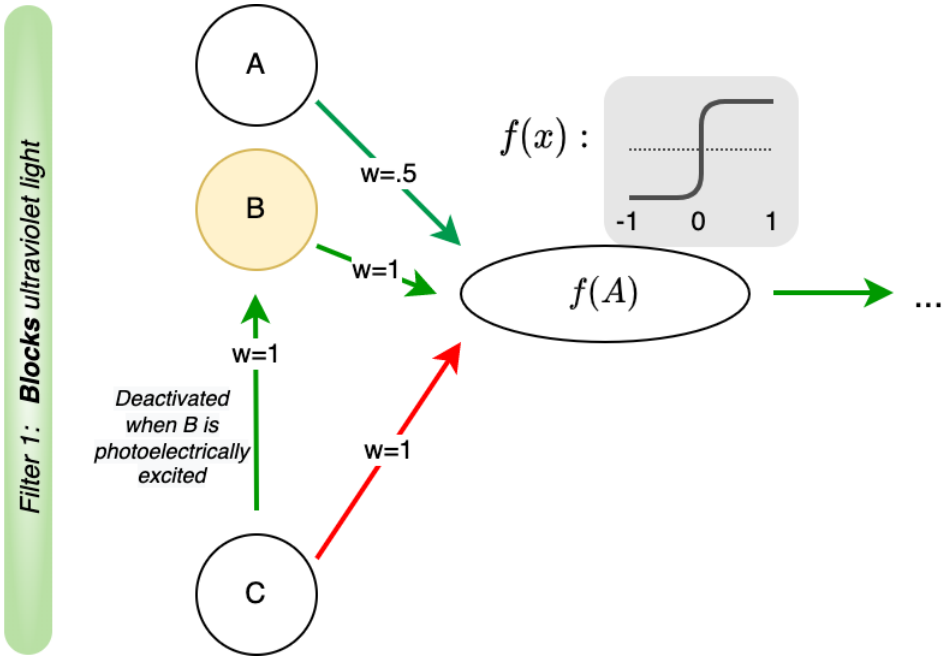

Figure 2.1: A network with a simple computational structure.

that this network exists in an organism with a cornea that (like ours) filters light before it reaches the input nodes. Specifically, in this organism, the cornea filters out all ultraviolet light.

The filtering means that B will always respond just as C does. And because B and C respond identically, the output cell appears to have a very simple job: it computes the threshold function for A, $f(A)$. Perhaps this is useful because the presence of A indicates prey, and the system can use this information further on. Regardless, B and C are irrelevant, cancelling each other out with the same signals of opposite polarity. The output node doesn't have any *computationally relevant* connections to B and C. These connections to B and C might be spandrels, a necessities for some other purpose — computational or otherwise — or whatever. But they are computationally irrelevant; they won't figure into our computational explanations of the system (or so I'm assuming for now). Note also that the *weights* of the connections are computationally irrelevant. If A's activity was being compared to a different node's, the computational description would have to take weighting into account. But because B and C cancel each other out, there is no more reason to describe the output node as computing $f(.5A)$ than there is to describe it as computing $f(.75A)$ or $f(3.827A)$. The weight of a connection tells you how strong it is relative to others, and if the

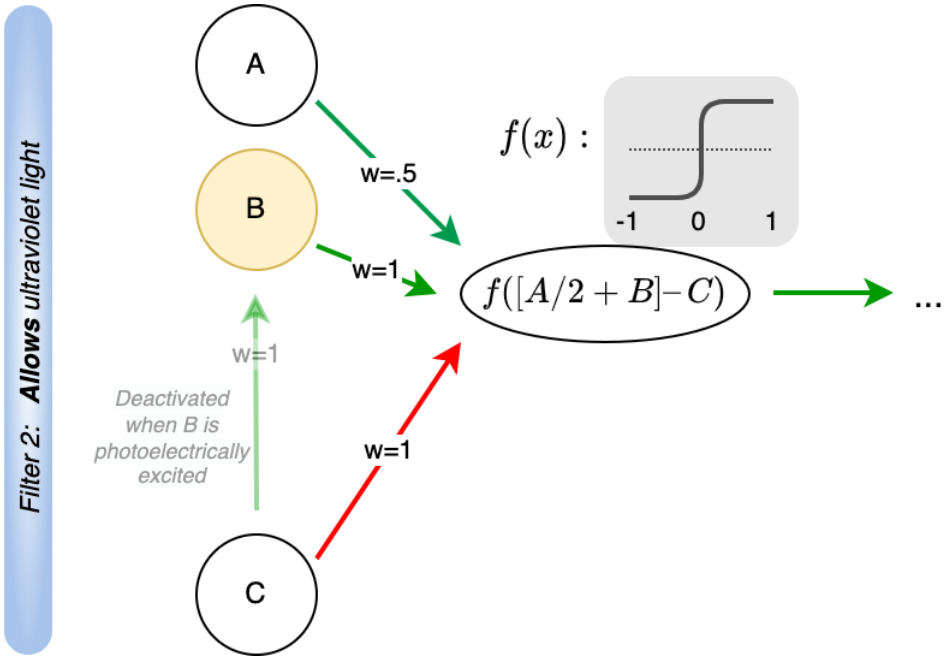

Figure 2.2: The same network as above, now with a different computational structure.

other connections are computationally irrelevant, then the weight of the remaining connection is too. All we're seeing so far is that many parts of the system's causal structure simply don't rise to the level of *computation*.

But now imagine this little system evolves a new cornea, which allows ultraviolet light through. The new network is illustrated in Figure 2.2. Then the connection between B and C will be severed in anything but the most artificial environments, and the output node will have a lot more information to work with. We haven't intervened *on the circuits themselves*, just on the cornea. Nothing about the circuits themselves has changed. But now it would clearly be wrong to say that the output node is just computing $f(A)$. B and C have independent effects on the node, and what it is really doing is computing the threshold function on some weighted combination of A, B, and C, or the weighted difference between A + B and C. You might imagine the new computation allows better tracks the same prey as $f(A)$ did, or tracks a wider range of prey. Regardless, we suddenly have a more complicated computational structure. And this isn't the only new computational element. Now it *is* misleading to leave out the difference in weight between A and the other two nodes. With the original cornea, the weight of A had no computational significance, no role in describ-

ing the system's computational structure. But now the weightings matter because A's activity is discounted by half before comparison with the activity of B and C. As new cells became computationally relevant, new aspects of the system's causal structure became computationally relevant, and the output node is now computing $f([A/2 + B] - C)$ rather than $f(A)$. The point of all this is just that the computational structure has changed, in the sense that *the causal structure that a computational explanation should capture has changed*, and that this has happened without any change to the computing circuits, just the cornea.

Now, to prove that these systems have their computational structure externalistically, we would have only to show that the basis or principle on which a re-description is required after the evolution of the new cornea does not just appeal to internal features of the system. We might argue, e.g., that the computational re-description is required only because the mutation allows the organism to track prey more effectively. Or we might, as I will in the case study to come, make the negative point that *the limited internal changes to the system are not sufficient to justify the computational re-description*. I will not be making that argument about the toy systems. *This is not the argument for causal structure externalism about computation.* That will come in the next section. This is just an illustration of what that externalism could look like, and a demonstration that causal structure externalism is not absurd.

It would be pointless to prove externalism about these systems because my ultimate target is externalism about *computation in cognitive science*, not externalism about computation in toy systems that bear a superficial resemblance to the ones cognitive science studies. And I probably *couldn't* prove externalism about these systems, because I couldn't prove that they really have the computational structures I've described. So much more information than I've given here goes into determining a system's computational structure. Computational structure isn't something you *just see* in a system's causal structure; it's something you *discover* about that system within the many and essential constraints of whatever field or inquiry your search takes place in. So the arguments will come when we're talking about the real systems that cognitive science studies, and about the constraints and demands on computational structure that cognitive science operates under. The

point of the toy systems is just to set us up to make sense of the arguments to come, and to make it clear what would make causal structure externalism about computation true. It would be true if the coarse-graining that ascriptions of computational structure involve must be motivated by, justified by, or grounded in environmental considerations — i.e., if internal considerations aren't enough to explain why, in one system, certain aspects of causal structure rise to the level of computation, while in another system they do not.

2.3 The argument for externalism

With arguments about my toy systems set to the side, the important take-away is an understanding of what the argument for externalism will look like, in the similar but real and more complicated case I'll discuss in this section. The argument will point out a change in the computational structure of a system, and argue that it is difficult or impossible to motivate or justify this change by appeal to taxonomic principles that refer only to the system's internal parts. That would mean the system's computational structure is determined partly by things external to it. An argument for externalism therefore has two main burdens: to show that there is a computational difference between two systems, and to show that this computational difference cannot be justified by reference only to the internal features of the systems. I'll do this with a case study from the cognitive neuroscience and evolutionary biology of color vision (section 2.3.1), which illustrates a system whose computational structure changed with extremely minimal change to its internal causal structure. Then (section 2.3.2) I'll argue that these minimal changes to its internal causal structure cannot themselves motivate the required computational re-description.

2.3.1 Case study: color vision

The finding in this section should be surprising, though it won't seem to support externalism on its own — the rest of the camel comes in section 2.3.2. The finding is this: the evolution from dichromacy to trichromacy likely occurred without any changes to post-retinal circuitry, but it calls for changes to the computational classification of post-retinal circuits. I'll start by discussing the

computations involved in trichromatic color vision. Then I'll summarize the relevant evolutionary biology and move on to the main finding. I'll finish by considering some possible resistance to the finding.

Humans, along with the rest of the old world primates, are behavioral trichromats, meaning we extract three dimensions of color experience from visual stimuli (Jacobs, 2002, 2009; Jacobs & Nathans, 2009). That is, we can find three monochromatic lights, i.e. three wavelengths, such that for any color we see, that color is indistinguishable from some weighted combination of those three wavelengths (Jacobs, 2018). For some colors less than three lights are required, but for some all three are needed. (The definition of dichromacy simply replaces "three" with "two.")[8]

There are a couple common misconceptions about this definition that I should clear up. First, trichromacy does not follow from the possession of three types of cone cell. More generally, n-chromacy does not follow from the possession of n types of cone (Jordan et al., 2010; M. Neitz & Neitz, 2014). Second, the three dimensions mentioned in the definition of trichromacy are not hue, saturation, and lightness. We distinguish hue alone on more than one dimension (e.g., see Derrington et al., 1984), and dichromat humans do not lack the ability to detect differences in hue, saturation, or lightness (Wachtler et al., 2004).

Moving on, the standard approach to color vision is summarized by Kelber and Jacobs as follows:

> Colour vision — the ability to discriminate spectral differences irrespective of variations in intensity — has two basic requirements: (1) photoreceptors with different spectral sensitivities, and (2) neural comparison of signals from these photoreceptors. (Kelber & Jacobs, 2016, p. 106)

Corresponding to (1), most organisms have a mosaic of cone cells in the first layer of their retinas, differing only in the opsin they express. Opsins are molecules that react to light, triggering a change in the membrane potential of their cell and causing it to send signals to post-receptoral

[8]The possible tetrachromacy of humans in mesopic light conditions (Zele & Cao, 2015), where rods are contributing to perception as well as cones, won't have any bearing on the following, so I leave it aside. Likewise for any possible tetrachromacy resulting from melanopsin-expressing ganglion cells (Horiguchi et al., 2012).

cells. Humans, and the rest of the old world primates, have three types of opsin, and thus three classes of cone. Step (1) in these primates involves cone cells expressing either an S, M, or L opsin, so named for their peak sensitivities to short-, medium-, and long-wavelength light, respectively. Most light sources stimulate each of these cells to some degree, but some mechanism has to collect these responses and determine a single color percept. That's step (2).

The mechanisms involved in step (2) are post-receptoral circuits that perform a few simple computations. One is

$$aL(\lambda) - bM(\lambda),$$

where $L(\lambda)$ and $M(\lambda)$ are the activity of L and M cones for a particular wavelength profile λ, and a and b are weightings of the L and M cone responses (Shevell & Martin, 2017). For simplicity I'll suppress the weights and wavelength-specifications and just call this the L – M computation. There also appear to be L + M and S – (L + M) computations in early vision. There is debate over many details, including the relative weightings of the L, M, and S terms in each computation and the adequacy of different sets of computations to the physiological and psychophysical data (J. Neitz & Neitz, 2011). But these debates are inessential for my purposes. What matters is that color vision is performed, in humans and our close relatives, by post-receptoral mechanisms implementing computations defined over the activation levels of the three types of photoreceptor — everything I say can be translated to different models of those computations.

These computations explain a lot about human color vision. They explain behavioral data to do with unique hues, color mixing, and opponent colors, among other phenomena (for two classic examples, see R. Devalois & Devalois, 1993; Hurvich & Jameson, 1957). To give just one example in more detail, the L – M computation defines what is know as the red–green opponent axis. When L – M is positive, a stimulus appears red, and when it is negative, the stimulus appears green. Really, this is only roughly true, but it is true enough that it is considered a significant explanatory success of the computational model, and an elegant explanation of the previously perplexing fact that a stimulus generally cannot appear red and green at once — reddish-green — though it can appear reddish-blue or reddish-yellow. The blue–yellow axis, S – (L + M), is partly independent of

the red–green one: L – M can be positive while S – (L + M) is positive or negative. But of course L – M cannot be positive while L – M is negative.[9]

The existence of these computations is also supported by physiological findings. They generate interesting hypotheses about the organization of the retina, many of which have been confirmed. One example is the conjecture and eventual discovery of midget ganglion cells with the right inhibitory and excitatory connections to cone cells to perform some of the computations I've mentioned (R. Devalois & Devalois, 1997). This is an ongoing project, with current efforts devoted largely to finding retinal ganglion or lateral geniculate nucleus cells with the right connections for the S – (L + M) computation (Dacey & Lee, 1994; J. Neitz & Neitz, 2017), or finding different computations that can be performed by cells with the organization we have found in the retina and lateral geniculate nucleus (Conway et al., 2010; Conway et al., 2018; Jacobs, 2014). Another ongoing project investigates the weights of L, M, and S cell input to the midget ganglion cells responsible for the computations, looking for principles that explain the relative uniformity of the midget ganglion response profiles despite the random process generating their connections to cone cells (Chang et al., 2013; Sabesan et al., 2016). The particulars of these last two paragraphs are not essential to the following; they are only an illustration of the depth and interest of this area of research, none of which would be possible without the computational model sketched above.

But evolutionary biology adds an important dimension to this story. We need a little more background first. In primates and other mammals the S opsin is encoded on a non-sex chromosome — one that every member of the species receives. The M and L opsins are encoded on the X chromosome, so males of the species get just one copy of those, while females get two. Our ancestors had only one of the longer-wavelength opsins, the M opsin. They were dichromats per the definition above, only able to extract two dimensions of color from stimuli (Jacobs, 2002, 2009; Jacobs & Nathans, 2009; Nathans, 1999). Males and females were in the same boat: males received one copy of the M opsin gene on their X chromosome; females received two copies, one on each X chromosome. Due to random X-inactivation, about half of the female's cone cells would express

[9]For more on opponent operations and their explanatory successes and shortcomings, see Shevell and Martin (2017).

one chromosome and half would express the other (Jacobs, 2008), but since both chromosomes had the same opsin gene the the result was a retina that looked just like the males', with only S and M cones.

The surprising finding in this literature is that the evolution from dichromacy to trichromacy seems to occur immediately on the introduction of a third type of photoreceptor cell via a transcription error affecting one of a female's two copies of the M opsin gene, resulting in a new allele: the L opsin gene (Dulai et al., 1999; Jacobs, 2009; Jacobs & Nathans, 2009). Now, because of random X-inactivation, her retina will now express the M opsin gene and the L opsin gene about equally, and suddenly she is a behavioral trichromat, without any change to her post-retinal circuitry. This is widely endorsed in the literature,[10] and it matters for my purposes because the computations we ascribe to the dichromat's and trichromat's early visual circuits are different. There is comparatively little work done on dichromacy, but what work there is does not presume or support the computational equivalence of dichromat and trichromat organisms. Perhaps more important is that the trichromat's post-receptoral computations are defined over three terms — S, M, and L — and the dichromat doesn't have any cells corresponding to L. In other words, the dichromat cannot be computing L – M because it doesn't have any L cells. It's not just that we would have to call the L cells something else; there is no class of cells that can be isolated or grouped together for the computations to be defined over. They're all M cells.

I'll revisit that point shortly, but let me first set out some of the direct evidence for this finding. Evidence from the similarity of the M and L opsin genes, along with their locations on the X chromosome, strongly suggests that one resulted from a transcription error copying the other (Dulai et al., 1999; Jacobs, 2009; Jacobs & Nathans, 2009). It took two steps to get where we are today. A transcription error resulted in an X chromosome with the new opsin gene instead of the old (introducing it into the population as an allele), and a later recombination error placed the two opsin genes next to each other on the same X chromosome (Dulai et al., 1999; Nathans, 1999). At

[10] See, among many other articles: Conway et al. (2010), Huberman and Niell (2011), Jacobs (2008, 2009), Jacobs and Nathans (2009), Kóbor et al. (2017), Mancuso et al. (2010), Mancuso et al. (2009), Mollon (1984), J. Neitz and Neitz (2011), M. Neitz and Neitz (2014), Shapley (2009), Wachtler et al. (2004), and Wachtler and Wehrhahn (2016).

the intermediate stage, any female who received both opsin genes — one on each X chromosome — would have a three-opsin retinal mosaic (because some cones would express one X chromosome and some cones would express the other). It is not until the second stage that males, with just one X chromosome, could have a three-opsin mosaic.

There are at least two reasons to think that a three-opsin mosaic would immediately result in a new dimension of color vision[11] (aside from the appeal to scientific consensus I made above). The first is the requirement of a selective advantage to spread the allele through the population (Jacobs & Nathans, 2009; Mancuso et al., 2010). Likely, that advantage would have been an ability to discriminate ripe fruit from foliage, or possibly skin or hair tone and socially important features of conspecifics (Dulai et al., 1999; Jacobs, 2009; Jacobs & Nathans, 2009). But because of familiar worries about adaptationism (Gould & Lewontin, 1979), I'm going to set this aside. The other reason is that the evolutionary step has been recapitulated experimentally in mice and new world primates. Mice have the same setup as our dichromat ancestors, with an S opsin gene on a non-sex chromosome and a longer-wavelength opsin gene on the X chromosome, without alleles. Jacobs and colleagues inserted a new long-wavelength opsin allele into knockout mice's X chromosomes, and at adulthood heterozygote females — females with both alleles, and thus a three-opsin cone mosaic — showed an extra dimension of color vision, discriminating between colors that their two-opsin conspecifics could not (Jacobs et al., 2007).

The second experiment was performed with adult male squirrel monkeys (Mancuso et al., 2009). Squirrel monkeys have the intermediate setup I mentioned above: there are different alleles of the X chromosome in their population, but there is still only one opsin gene per X chromosome.[12] Male squirrel monkeys are therefore dichromats, and females who are lucky enough to

[11] I'm going to treat this phrase, "a new dimension of color vision", as substitutable with "trichromacy". Some authors are reluctant to equate the two — e.g., Gerald Jacobs (personal communication), who performed one of the experiments to follow. I think he is reluctant to equate the two because in his experiments on mice, the new dimension of color vision is confined to a portion of the wavelength spectrum on which, previously, it's possible that only one wavelength was needed to match any percept — i.e., it's possible that the mice were monochromats with respect to that section of the wavelength spectrum before the mutation, and dichromats after. But it is nonetheless true that the new cone type increased the dimensionality of color vision along one portion of the wavelength spectrum without altering the post-receptoral circuits involved; that scenario shouldn't look very different from the one where across the board dichromacy is turned into across the board trichromacy, which is the situation I'll discuss.

[12] Squirrel monkeys have this in common with the rest of the new world monkeys, the one exception being the

receive different alleles on their two X chromosomes are trichromats. The experimenters injected a virus into the male monkeys' retinas carrying an allele of the M opsin, along with genetic instructions to express it. Very shortly, when the new opsin was expressed in a significant number of cones, the monkeys became trichromats.

So in both mice and squirrel monkeys, the finding holds up: a new type of photoreceptor, with no changes to the circuits responsible for color vision, is sufficient to transform a dichromat into a trichromat.[13] And as I said above, the move from dichromacy to trichromacy calls for a computational re-description of the unchanged circuits. This result is not yet a challenge to internalism: we're talking about an internal change leading to a change in computational structure. This is where the definition of externalism, which took up so much time earlier, is important. On the supervenience definition, this could not be evidence for externalism: the change in computational structure came along with a change in internal structure. But on the "reference to the environment" definition, the question is whether a taxonomic principle making reference only to internal features of the organisms can *support* or *ground* a computational re-description. I'll argue against that in the next section.

But first, I've emphasized that this result is quite surprising. Why would a two-opsin dichromat already have the setup required to take advantage of a third type of opsin? It's not just that the mechanisms are number-of-cone-types-general, otherwise there would be many human tetrachromats and pentachromats — humans have various alleles for the M and L opsin genes, and a significant percentage of women have and express distinct alleles for one or both on their two X chromosomes despite being trichromats (Jordan et al., 2010; Jordan & Mollon, 2019). Since the finding is so surprising, I'll spend some more time on it before turning to the alternative. Specifically, I'll discuss two potential ways of resisting the finding.

My point has been that the evolution from dichromacy to trichromacy likely occurred without any changes to post-retinal circuitry, but it calls for changes to the computational classification of

howler monkey, which is a trichromat that evolved to have two opsin genes per X chromosome, very likely by the same evolutionary process as the one I've described (Jacobs, 2002, 2009).

[13] Possible changes to the circuits due to plasticity are discussed shortly.

post-retinal circuits. So there are two main ways to dispute the point. Maybe trichromacy *doesn't* actually call for changes to the computational classification of post-retinal circuits — I'll discuss that first. Or maybe there *were* changes to post-retinal circuitry — I'll discuss that momentarily.[14]

Objection 1: there is really no computational re-description required. For this to make sense I would have to have been wrong about either the dichromat's or the trichromat's computational structure, or both. Other models of trichromacy don't tend to classify trichromats and dichromats as having the same computational structure, so it's hard to support this objection with scientific evidence. Instead, it tends to be grounded in philosophical concerns. But I pointed out above that the trichromat's computations are mysterious in a dichromat. Of course, one can gerrymander a system however one likes — if all the same circuits are there, we could say that a circuit computing L – M in a trichromat was originally performing the same subtraction in its dichromat predecessor, just over different cells. But the supposed subtraction in an organism without L cones is just the function M – M, a constant function to 0. It is hard to justify ascribing that computation to the system, since it appears to do nothing. Compare the toy example from earlier. It would be hard to justify attributing to the original system a C – B computation, let alone the one that is more analogous to this case: a C – C computation.[15] And there is ongoing research investigating how trichromatic color vision could have been supported by visual circuits that had not developed to support it. Some researchers suggest that trichromacy is supported by mechanisms that originally

[14]I'll also mention here a small group of dissenters who argue that, at least in the experiment with mice, the animals did not develop trichromacy, but merely new sensitivities to texture and illumination that can be mistaken for trichromacy (Cornelissen & Brenner, 2015; Makous, 2007). This is less plausible in the primate case, though, and for lack of space I will focus on the more plausible and common objections.

[15]Though not as hard as it might initially seem. The following example is due to John Morrison. You can imagine a neuron wired up randomly to M cones, half of its connections excitatory and half inhibitory, so that it effectively subtracts the activity of a random set of M cones from the activity of another random set. One use for this might be to measure the *variance* among M cone activity. This could be a useful operation, and it's well-described by the computation in question: a simple subtraction that computes M – M in a dichromat. The point is that the mechanism computing M – M in a dichromat might do a decent job of computing L – M in a trichromat, as long as its excitatory and inhibitory connections were each dominated by different cone types. So I don't want to make an in-principle argument against M – M computations altogether. But I don't know anyone proposing this computation, or a similar one, as a model of what the relevant cells were doing pre-trichromacy. I set out the mainstream proposals in the next few sentences, and neither supports computational equivalence. So while it is not impossible in principle that the circuits in question are computing M – M, it is implausible in this case. And that's all I need for the argument that the post-retinal circuits do, in this case, have a new computational structure.

served spatial vision (Shapley, 2009; Wachtler & Wehrhahn, 2016);[16] others suggest models of dichromacy involving mechanisms that would perform different color vision-related computations in dichromat and trichromat retinas (J. Neitz & Neitz, 2011), and so on. We should not constrain this research by an assumption that the relevant mechanisms in dichromats must perform the same computation they do in trichromats.

Another way to make this objection is to say that, although the trichromat and not the dichromat is 'computing' L – M, that's not *really* their computational structure. One version of this view might say that their real computational structure is at a lower level of description, e.g. at the level of individual midget ganglion cells and their particular sets of synapses onto particular cone cells. Those haven't changed between the two organisms;[17] just the opsin expressed in the relevant cone cells has changed. So perhaps only a *higher-level gloss* of the computational structure has changed; the underlying computational structure hasn't. This objection is the one I most frequently hear (from philosophers, at least), and the one that is most deeply mistaken. The objection amounts to a flat denial that the trichromats in question are implementing the computations that are attributed to them to explain their trichromacy. What they're really computing is something else. But to retreat to a 'different level' of 'real' computational structure is just to ignore the topic at hand, which is the way computation works *in cognitive science*. Unless you are willing to describe *every* causal structure as a computational structure — making your view especially irrelevant to computation in cognitive science — you will need some way of saying which structures are computational and which aren't. But you've given up the main constraint, which was to understand cognitive scientific practice, and are now making decisions about what counts as computational *in spite of* cognitive science. This makes it difficult, if not impossible, to say anything informative about computation in cognitive science. What matters is that, in cognitive science, the computational

[16]That mechanism might look a lot like the M – M one: a center-surround opponent cell might be useful for spatial vision, and might also perform an L – M computation if its center and surround contained different densities of L and M cells. But insofar as it serves spatial vision it does so via a center–surround computation — not by computing M – M (see Wachtler & Wehrhahn, 2016).

[17]At least we can stipulate that they don't. All along it has been an idealization to assume that two organisms have *the same* post-receptoral circuits, because those circuits are generated stochastically and will differ between any two organisms. But the important point — a point this objection accepts to get at something deeper — is that the post-receptoral circuits don't differ in a way *specifically related to trichromacy*.

explanations of color vision are the ones that attribute computations like L – M to trichromats. The pre- and post-mutation organisms do have their lower-level causal structure in common, but as far as cognitive science is concerned that simply isn't the level of description, or at least it isn't the *only* level of description, where computation is going on. It's really (or also) going on at the level of description where your post-receptoral circuits compute L – M and things like that. So it is only on a revisionary and irrelevant sense of 'computational structure' that this objection can be maintained.

Objection 2: a computational redescription is required, but only because there were *changes to post-retinal circuits, likely due to plasticity.* One might accept that the post-mutation organism has a different computational structure than her predecessor, but think this is because her post-receptoral circuits developed differently (Nathans, 1999; Wachtler et al., 2004). After all, the mice had their entire adolescent period to take advantage of the new features of their photoreceptor layer — is it not conceivable that their post-receptoral mechanisms developed to take advantage of that photoreceptor layer? This is tempting, but look back to the experiment on squirrel monkeys: it was performed long after critical development periods — when the brain is most plastic — had ended (Feldmann et al., 2018; Hubel & Wiesel, 1970), and the monkeys developed trichromacy "just as levels of transgene expression [the presence of the new opsin] became robust" (M. Neitz & Neitz, 2014). The experimenters argue that this is much too soon for the new visual capacity to be due to plasticity in early — especially retinal — circuits, where the computations at issue are performed (Conway et al., 2010; Mancuso et al., 2010; Mancuso et al., 2009). So this objection needs to take a more complicated route.

What if we divide and conquer the two cases? We could say that the initial evolution of trichromacy — in the mouse experiment and in the evolutionary history of squirrel monkeys — was due to plasticity. That would explain the mouse experiments just fine (Jacobs & Nathans, 2007). Then we could say that in species like squirrel monkeys, where the females are already trichromats, there was some selective pressure towards a *hardwired* trichromatic setup. The first female owners of a third cone type would have had to develop trichromacy through plasticity like the mice, but

then a hard-wired trichromatic visual system could have developed over evolutionary time for the benefit of these females, and that hard-wired visual system would explain the speedy development of trichromacy in later males. We would have to assume that developing trichromacy through plasticity is less effective or efficient than having a hardwired system, and that the benefit to some fraction of females of more efficient or effective trichromacy was sufficient evolutionary pressure for the hard-wired system to develop as a species trait, despite being no benefit or even a detriment to male squirrel monkeys (given that the hardwired system forces them to compute, among other things, a constant function to 0 rather than something more useful). Assuming that both assumptions are reasonable, this would mean that the squirrel monkey experiment couldn't be called in to defuse the plasticity worry that was raised in response to the mouse experiment. The mice developed trichromacy via plasticity, so there were not common circuits between the di and trichromats in their case. And the monkeys started with hardwired trichromatic circuits, so they are not an example of previously dichromatic circuits supporting trichromatic vision.

So we have two interpretations on the table. They raise an unfortunate number of issues at the very coalface of neuroscience, especially concerning plasticity. Needless to say, I can't go into enough detail, in a paper like this, to to conclusively refute the alternate interpretation. And it's not likely that there is anything conclusive to say right now, since this is an area of ongoing research. But there are three reasons to be suspicious of the second interpretation, especially in this context.

First, consider *currently* dichromatic male squirrel monkeys. The point of the plasticity objection is to convince us that they have the same computational structure as their female or genetically-altered male conspecifics. But this runs into a familiar problem: we end up attributing to them computations like M – M that don't figure into our models of dichromacy. If we want to avoid that result (and I've suggested that there is good reason to), then, regardless of plasticity's role in all this, we end up at the same place. Because we want to count the dichromat male squirrel monkeys as computationally different from the trichromat ones, and because we allow that the male squirrel monkeys develop trichromacy without the need of plasticity, we end up saying that which computational structure a post-receptoral mechanism counts as having depends on differences in the

retinal mosaic alone, not on post-receptoral circuits.

Second, if trichromacy results from a plastic reorganization of post-receptoral circuits, there seems to be no reason that plasticity would be able to accommodate three types of cone cells but not four or five, and so the many human females with four or five cone cells should be tetra- or pentachromatic. But this is not the case, and the evidence that even *some* human females are even *tetra*chromatic is controversial and highly inconclusive (Jordan et al., 2010; Jordan & Mollon, 2019).

Finally, not every instance of neural plasticity calls for a computational re-description. The brain is changing constantly due to plasticity, but we do not need new computational explanations of it from second to second as synapses are added or destroyed. That means that the challenge I pose in the section 2.3.2 can be raised *even given* the plasticity objection. Perhaps it can be met more easily if we grant the plasticity objection, but that remains to be seen. Anyone partial to this objection should therefore revisit it after section 2.3.2. The argument there goes more smoothly if we assume that the post-receptoral mechanisms haven't changed, but I don't see a clear way forward on the challenge, from an internalist's point of view, even if we grant that they *have* changed.

2.3.2 Finally, the challenge

So far we've seen evidence that the computational identity of the neural circuits responsible for color vision depends on the make-up of the photoreceptor mosaic. This doesn't establish externalism, but it does pose a challenge that an internalist might have trouble meeting: to find a plausible taxonomic principle that grounds the computational difference between the pre- and post-mutation organisms — call them Pre-M and Post-M — given the extremely minor changes to their causal structure.[18]

The externalist has some options, of course, though I won't discuss them in detail. E.g., we might distinguish between Pre-M and Post-M partly on the basis of their differing abilities to ex-

[18]Or, if you are partial to the plasticity objection, given a more significant change to their causal structure. But, as I've indicated, I'll be setting this aside.

ploit certain features of their environments: Post-M can take advantage of differences between stimuli that Pre-M can't take advantage of, and a computational explanation should explain how she does this. If this is what motivates computational re-description, we have the start of an externalist account: facts about the organism's possible interactions with her environment constrain the ascription of computational structure. This is plausible, if sketchy. It is, however, a plausible sketchy externalist principle, and I want to leave the externalist options aside (though see Richmond, n.d.-c, for a start on an externalist view). The purpose of this section is to sow pessimism about the *internalist's* ability to meet the challenge.

So what can the internalist say to ground the computational difference between Pre-M and Post-M? Consider some poor first attempts, just for the sake of illustration. It can't be merely that there is *some* internal difference between the two organisms, nor that there is *the specific internal difference there is* between Pre-M's and Post-M's retinas. The former overgenerates distinctions. E.g., it would distinguish Pre-M's computational structure from Post-M's, but it would also distinguish Pre-M from a silicon duplicate, or a duplicate with just one neuron replaced by a silicon chip — paradigmatic examples of a non-computational difference. The latter under-generates distinctions because it is ad hoc — it fails to serve as a taxonomic principle or a *basis* on which taxonomic decisions are made. The main constraint that these first passes fail to meet is that our taxonomic principle should make it clear why the computational re-description is explanatory, fruitful, or correct, or at least tell us with some generality when re-description is called for.

What more serious possibilities are there? A taxonomic principle distinguishing Pre-M and Post-M must appeal to their differences, and for the internalist these could only be (i) a difference in their photoreceptor mosaics, or (ii) a difference in some other internal feature — not the circuitry, since that's the same, but perhaps the patterns of activity in neural circuits or in the organism more broadly.

(i) would cause the internalist no end of trouble. As we saw, the fact that there is *some* difference between Pre-M and Post-M is no help to taxonomy, nor is the *precise* difference between Pre-M and Post-M. But nothing in between seems to have much promise either. The specific in-

volvement of a new type of cone — L rather than M — is no help, because replacing all M cones by L cones would have resulted in no computational re-description, as is clear from the existence of alleles in new world primate populations: dichromat males can differ in which allele they receive, but aren't understood to be computationally different as a result (Jacobs, 2008). The same goes for human trichromats with anomalous alleles of the M or L opsin.

The fact that three types of opsin rather than two are now involved is no help either, because the third opsin might have been an M_2 opsin with the same light sensitivity as the original M opsin, which calls for no computational re-description (that case would be analogous to a silicone substitution). It might also have been an opsin that had the same sensitivity but worked slower or faster, likely calling for no computational re-description despite a more consequential change.[19]

But perhaps this points the way to a better principle. Isn't what's important really the compound fact that there is a new type of opsin *and* that each opsin has a unique wavelength sensitivity profile? The move from three opsins to four in humans is reason to suspect that this alone isn't enough to justify a change in computational identity: many human women have four or five types of cone cell (Jordan et al., 2010) despite being trichromats, and not needing a unique computational explanation of their color vision. Many other organisms have more cone types than dimensions of color vision (M. Neitz & Neitz, 2014) as well.

It seems to be only when there is an increase in the dimensionality of vision (e.g. from dichromacy to trichromacy) that re-description is required. Can the internalist add to the compound fact a third component? There is a new type of opsin, each opsin has a unique wavelength sensitivity profile, *and* the organism becomes a trichromat. The fact that Post-M became a trichromat and human females with extra opsins don't generally become tetra or pentachromats is important, and may be what motivates a new computational description in one case and not the other. But this is no help to the internalist unless the difference between dichromacy and trichromacy can be cashed out in internalist terms. The standard definition (see the beginning of section 2.3.1) appeals to environmental light sources and the way an organism interacts with them. It's unclear what the

[19]See Shevell and Martin (2017) on some interesting effects of the speed and efficiency of the different types of signals cone cells send, which don't imply any changes to the computational structures I outlined above.

difference between dichromacy and trichromacy would be on internalist terms. The internalist's only resources are precisely (i) and (ii), above: features of the causal structure of the organisms — which I've just argued do not do a good job of distinguishing between dichromats and trichromats — or internal patterns of activity in that causal structure — which I'm *about* to argue don't do a good job. If these arguments are any good, appealing specifically to a change in the dimensionality of color vision is not a good option for internalists.

I'm not sure what else an internalist might say about (i), but I can't think of anything that doesn't obviously over-generate or under-generate. (ii) might look more plausible. Maybe what makes the difference between Pre-M and Post-M is something to do with their *patterns of neural activity*. The internal organization of the organism — its "causal topology" (Chalmers, 2011) — hasn't changed except in the retina, but within and beyond the retina the patterns of activity will change. The organism may have certain cells excited more or less often, certain circuits' activity more or less correlated with others, and so on. That would count as an internal difference, not an external one, and it is consistent with there being no change to any post-retinal circuitry.

But, of course, not just any difference in patterns of neural activity makes a difference to computational structure. Cases of visual deprivation establish this: something external to an organism can block stimuli (e.g. an eye patch) or modify them for significant amounts of time (e.g. reversing glasses; Harris, 1965), causing significant changes in the patterns of activity the nervous system undergoes (Miyauchi et al., 2004). But it's not clear that we should be changing our computational description of the organism on those grounds.[20] Recall the discussion in section 2.2.2: we had better not need new a description of an organism's computational structure every time there is some mundane change to its environment, like an eye patch coming off. So we need a well-motivated similarity metric or equivalence class that groups Pre-M and Post-M separately on the basis of their neural activity. Of course, it should also group Post-M *together* with, e.g., Post-M Anomalous: a trichromat with different alleles of the M and/or L opsin that have slightly different wavelength sensitivities, or different processing speeds, or so on (an extremely common phenomenon in hu-

[20]Of course, the actual computational operations the system performs performs may change *in response to* deprivation or reversal (e.g., see Hubel & Wiesel, 1970). But, again, the question is about *classification*, not causation.

mans and other species that doesn't call for unique computational descriptions; see Jacobs, 2008). Maybe there is some similarity metric that gives this result, and that can be non-arbitrarily justified. But I don't see any reason for optimism, and the burden is on the internalist to devise one.

Leaving aside the neural activity, what about changes in the organisms' *patterns of behaviour*, construed internalistically as, e.g., the totality of the movements of the parts of an organism. An internalist definition of behavior will have its troubles — see the discussion in Krakauer et al. (2017, Box 1). But even if we accept it, there is a difficult job to do explaining what differences in behavior so construed count as computational differences. Even stating a significant difference between the internalistically construed behavior of Pre-M and Post-M is a challenge. And then we must again explain why that is computationally significant, but the difference in behavior between Post-M and Post-M Anomalous is not. This has all the same challenges I noted in the last paragraph.

I can't say anything more conclusive here. Internalism is not self-contradictory, and it's unlikely that it will be outright inconsistent with observed phenomena since it doesn't make empirical predictions but merely guides the construction of theories, directing attention to one set of facts rather than another. Rather than a contradiction or false prediction, what should convince us to reject internalism are situations where paying attention to internal features just doesn't pay off. In the case at hand, those are situations where the taxonomy we need isn't generated from internal features. I've argued that the case of color vision presents just such a situation. The facts available to the internalist — changes to the cone mosaic and post-receptoral patterns of internal activity — offer no clear way to make the taxonomic distinctions we need. I haven't shown that it is *impossible* to find a route from those facts to those distinctions, but this is a problem that any internalist account has to solve. The point of this section was that we should be skeptical of their ability to solve it.

OK. Challenge posed; pessimism sown; internalism argued-against. It would be premature, at this point, to move directly to a particular version of externalism, e.g., one on which the brain's computational structure is determined by its representational contents or its teleological properties.

An externalist account along those lines would have to be derived from the way that the environment supports and constrains computational explanations. There is much work left to do for such an account, which I won't undertake here (though see Richmond, n.d.-c). My argument has simply been against internalism, and so for *externalism simpliciter*. But even the bare thesis of externalism provides a way into important cognitive scientific debates, and I want to conclude by discussing some of them.

2.4 Externalist methodology in cognitive science

In section 2.2, I drew on the debate over behavior and ethology in neuroscience to illustrate the relevance of externalism to cognitive science. I want to revisit that issue now, with a few examples, to give a sense of the interventions an externalist can make in current cognitive science.

Krakauer et al. (2017) raise problems for an internalist approach — one that takes an organism's computational structure to be independent of its environment. They describe this as the question of whether "neuroscience needs behavior," with behavior understood externalistically in terms of its interactions with the environment, not just as the total movements of an organism. As I mentioned earlier, these authors, and cognitive scientists generally, understand this as a methodological question rather than a substantive one. But the externalism I've argued for bears on the methodological question. As Krakauer et al. (2017) put it, in an analogy with computer science: neuroscience seeks to understand the (computational) processes governing behavior, and the "core question ... is whether the processes governing behavior are best inferred from examination of the processors" (p. 480). On the analogy, the processors are neurons and neural structures. The answer, if I'm right about externalism, is *no*. The processes governing behavior are not best inferred from examination of the processors, because they cannot be inferred *at all* from examination of the processors. They depend not just on the processors but also on the organism's environment, and any inference to an organism's computational processes or structure needs to consider the parts of the environment that contribute to determining that structure. So if I'm right about externalism, "neuroscience needs behavior" in something like the sense that Krakauer et al. argue.

What would an alternative, internalist approach look like in cognitive science? Connectomics is a great example. Connectomics attempts to understand the brain's structure by mapping every neuron in the brain, and every one of its synapses with other neurons. We have good connectome maps of certain simple organisms like C. elegans, and we have a promising start on small sections of the human brain. There is a growing appreciation in neuroscience for the power of this approach. And while that appreciation is not entirely misplaced, if externalism is true it must be tempered significantly. To understand an organism's behavior it will undoubtedly be necessary to understand its fine-grained neural organization, at least in some respects and to some degree of approximation. Even theorists who see computational modeling as radically autonomous from neuroscience (Gallistel & King, 2009) agree that questions of computational implementation are answered by neurological details. But the appreciation of connectomics is misguided insofar as it derives from a belief that the connectome will itself settle questions about the computational structure of the brain (Schneider, 2019, p. 115), or will settle those questions with relatively little input from our understanding of the brain's environment (Seung, 2012). That is in straightforward contradiction to the conclusions of the previous section. Only in conjunction with the environment does the connectome determine a system's computational structure. The causal structure of the brain, insofar as it figures into computational explanations or is captured by models of the brain's causal structure, is simply not something the brain has in itself. It is not something that a connectome map, no matter how detailed or accurate, can reveal.

It's worth noting that this is distinct from the common (and well-taken) criticism that connectome maps miss *non-neural* aspects of the brain's causal structure like glial cells and volume transmission (Anderson, 2014, Interlude 2). And it is distinct from the also common (and also well-taken) criticism that connectome maps have little to say about the *function* of neural circuits because they don't indicate the strength of connections, whether synapses are inhibitory or excitatory, and so on (Morgan & Lichtman, 2013). The point here is that even if we had all those things, connectomics alone would be unable to tell us the brain's computational structure. It is often pointed out that even our highly detailed connectome maps of C. elegans provide scant un-

derstanding of behavior (Niv, 2020, p. 134), and externalism explains why in a deeper way than these other criticisms: we look for the computational structure of the brain because (whatever it is derived from) it's the computational structure that explains behavior. But if computational externalism is true, we can't explain behavior just by looking at a connectome map, because that *isn't where the computational structure is to be found*.

Another place where externalists can make useful interventions is in the debate over the importance of naturalistic or ecologically valid stimuli. The externalism reached above doesn't, itself, settle that debate. The computational structure of an organism depends on its environment, but it is only together with some reason to care about a particular aspect of the environment that this can motivate a position on naturalistic stimuli. Here's an example. There is some interesting work on stereoscopic depth estimation by naive subjects as compared to repeat performers of an experiment (Hartle & Wilcox, 2016). In initial experiments, naive subjects overestimate short distances between objects and underestimate long distances, but subjects who sit through the experiment multiple times become quite accurate about both. This is surprising, because the repeat performers are not unique in training stereopsis — we're all using it, all the time. So what would make repeat performers so much better than naive ones? The experiment was originally done with images on a computer screen, designed to provide specific stereoscopic cues. But when the experimenters used real objects instead, the disparity between naive subjects and practiced ones disappeared. This likely means that the computer images had inappropriate *monocular*, i.e. non-stereoscopic, cues. The repeat performers learned to tune out the monocular cues that a naive subject couldn't help making use of.

So, which environment, which experimental condition, is informative about the brain's computational structure? Should we be proposing computations that take into account the repeat subjects' especially good performance on artificial stimuli as compared to naive subjects'? Or should we be proposing computations that take into account the shared accuracy of naive and repeat subjects on natural stimuli? With the artificial stimuli, repeat performers were ignoring monocular cues that are usually, and in natural contexts, helpful, and which are apparently taken advantage of

automatically and unconsciously in the absence of highly artificial negative reinforcement. That makes it sound like the they were doing things wrong, or at least *unnaturally*, in which case we should presumably set them aside when we think about the brain's computational structure. In that case we would propose a computational theory that does not artificially separate monocular and binocular cues (as the repeat performers did). But the repeat performers tuned out all cues other than stereoscopic ones; doesn't that mean their performance properly isolates one aspect of depth perception, stereopsis, that should come in for computational explanation? Then the natural environment should be set aside, and we should propose computational theories that take artificial stimuli and environments more seriously.

Externalism itself doesn't support one side or the other. You can hold that the computational structure of a system depends on its environment while holding any of the following three positions: that what really matters for determining computational structure is (i) stimuli that isolate the most fine-grained functions possible, (ii) natural stimuli, or (iii) both. There are ways of harmonizing the positions. E.g., perhaps the ultimate explanandum for cognitive neuroscience is how brains process naturalistic cues, but investigating the way they process artificially isolated subsets of those cues is a way of getting at aspects of the naturalistic process. But this is just one of many possible positions, and it is by developing the externalist thesis in more detail that philosophers stand to intervene on the debate, saying which aspects of the environment help determine the brain's computational structure and how.

Both these issues, about the fine-grained causal structure captured by approaches like connectomics and about naturalistic stimuli, arise often in cognitive science. To illustrate both in the case of color vision, consider recent technology that makes it possible to optically stimulate just one cone cell at a time and measure the responses of further cells (Sabesan et al., 2016). This is highly useful for understanding implementation details (like connectomics), but it is often taken to also reveal the computational structure of the early visual system (Kling et al., 2019). As I argued in the case of connectomics, if externalism is true, that kind of inference from physiological structure to computational structure is far too quick. Single-cone stimulation is also highly unnatural:

because of optical blur imposed by the cornea, outside the laboratory no scene is ever represented in the retina at single-cone resolution (Kling et al., 2019). To understand the significance of this research we need to know whether non-naturalistic stimuli help determine computational structure or not. So externalism simpliciter rules out an inference to computational structure from results in single-cone stimulation studies, and the further development of externalism would reveal the true significance of those studies.

There is much more to say about the relationship between externalism and cognitive scientific practice. The point in this final section has been that externalism, properly understood, settles some of cognitive science's questions and provides a starting-point to answer others. Continued development of externalism would answer more of those questions and bring the philosophy of cognitive science into closer connection with the cognitive sciences and their explanatory practices. In fact, aside from the argument for computational externalism, that has been the main upshot of this paper: an approach to computation that takes into account the way computational explanation works in cognitive science, and especially one that understands its philosophical questions as ones *shared* with cognitive science, can bridge some of the gaps between work in philosophy and cognitive science — a consequence both fields should embrace.

Chapter 3: What is a Theory of Neural Representation For?

3.1 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 & Tsao, 2017), supports language use by representing word meanings (Borghesani & Piazza, 2017), and so on. So one 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?

But, although this is a central question in the philosophy of neuroscience, it is not a *basic* or *fundamental* one. It is an important question only because the notion of representation figures into cognitive scientific explanations. The more fundamental questions are about how those explanations work, and why they work — why they are so successful. It is in pursuit of these questions that the problem of neural representation arises.

I'm going to propose a new way of answering questions about representational explanation: by illuminating the role that *representational notions themselves* play in the explanatory economy of cognitive science and especially neuroscience — what they make it possible or feasible for scientists to do, and how. I'll stress the way representational notions help us construct and understand models of the brain's causal structure. This will allow me to answer our questions about how and why representational explanations work with no need to define, or even commit to the existence of, a property of representation. This is not an argument against that property's existence, but

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

against its relevance for our projects as philosophers of neuroscience — a sort of *methodological nominalism*. I'll start with some examples of representational explanation in section 3.2. I'll then illustrate the standard philosophical approach to representational explanation in section 3.3, noting especially the emphasis it puts on a metaphysics of representation. In section 3.4 I'll outline an alternative, non-metaphysical approach, and in section 3.5 I'll use that approach to build an account of representational explanation, before discussing some objections in section 3.6 and concluding.

3.2 Examples: place cells and the fusiform face area

Many organisms have a remarkable capacity to navigate their environments, finding their way around obstacles, making their way to remembered destinations, and getting home from new locations along efficient paths. Our current understanding of how the brain supports spatial navigation started to come together in the 1970s 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, 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 & 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 & 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 & Yovel, 2006). There is debate over *how* the FFA individuates faces, but an interesting suggestion is that it does this by representing the precise way that different faces deviate from a "norm or average face" (Kanwisher & Yovel, 2006).

Neither case is uncontroversial. But the explanations of face perception and spatial navigation are shot through with representational notions, and it would be exceedingly difficult to get rid of them even if we wanted to. What do these representational 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, carries information about, etc., 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 & McNaughton, 2006). FFA activity is also famously correlated with many things aside from faces (Rhodes et al., 2004) and, to be fair, there is a case to be made that the FFA is better understood as representing features that are not at all unique to faces (Kasper et al., n.d.). But what's important, for my purposes, 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 (though these are common and useful targets for experimentation). 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.

3.3 Questions about representation

Neuroscientists ask various questions about representations. What feature of the environment does some neural activity represent? What computations does it perform on those representations? 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 so successful?

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. It aims to say which relationships between brain and environment[2] are representational, and why. As an answer to the philosopher's questions, this approach assumes that representational explanations work by attributing to neural activity a property — BEING A REPRESENTATION, or BEING A REPRESENTATION OF X — which is responsible for the behavior we're interested in explaining.[3] And it assumes that the way to understand representational explanation is to investigate this property it attributes to neural activity. If it is by referring to this property that representational explanations function, and if it is by accurately picking out instances of that property that they succeed, a successful metaphysics of that property could shed light on how and why representational explanation works. The neuroscientific questions would be answered in a similar fashion. A given neural structure or a bit of neural activity will either satisfy the definition of the property REPRESENTATION (and REPRESENTATION OF X) or not. That will tell us which neural structures or activities represent, what they represent, which parts of the brain implement the representations, and so on.

This approach is exemplified in classic work on representation, especially work that focuses largely on neural representation, e.g., by Cummins:

> Empirical theories of cognition can and do take the notion of mental content as an

[2]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.

[3]Ramsey (2007) and Mollo (2020) are particularly explicit about the standard approach's commitments.

> 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.[4] (Cummins, 1991, p. 2, emphasis mine)

And it is dominant in recent work as well. E.g., Shea, focusing like Cummins not just on mental representation but on neural representation in particular (Shea, 2018, p. 6), moves directly from the existence of representational explanations to puzzles about the property of representation 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)

Of course, I plan to doubt that lacuna. But it's worth noting three things before moving on. The first is this: the standard approach is broader than the project of defining representation 'naturalistically.' The philosophers I've cited tend to assume that the metaphysics of neural representation should be framed in causal, mathematical, biological, or other 'basic,' non-intentional terms. But you could take the standard approach without any such assumption. Many philosophers would argue that there are psychological representations grounded in neural states, and that the neural states count in an important and legitimate sense as representations, but would not be concerned to define these representations in basic or "naturalistic" terms. Consider Burge, who, in a different context, explicitly aims to give a non-naturalistic definition of *psychological* representation (Burge,

[4] 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, 1983, p. 2) — namely, "(actual) neurophysiological states" (Cummins, 1983, p. 6).

2010a, p. 296). There is, for Burge, no need to reduce the property of representation to something more basic, or something that is itself "not mentalistic" (Burge, 2010a, p. 296). I don't know of anyone with a similar approach to *neural* representation specifically, but one could imagine similarly non-naturalistic approaches. As an extreme example, perhaps a neural representation is a neural state with some specific relation to Cartesian mental substance. The point is that this view is still committed to the standard approach as long as it claims there is *something it is* for a bit of neural activity to represent, and wants to understand representational explanation by defining *that something*.[5] It is this approach I'm targeting. Not the naturalistic reduction of representation, but the broader view that our understanding of representational explanation in cognitive science, and especially neuroscience, should issue from an understanding of what neural representation itself is, of what it is for a bit of neural activity to represent.

Second, 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 largely in terms of the role that cognitive science needs representations to play, or the "job description" that neuroscientists set for representations, and which a bit of neural activity has to satisfy to count as a representation (24-25). But as that description makes clear, Ramsey still thinks that to understand representational explanation we need to investigate this 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, p. 188). Even more explicitly, he describes his approach as the analysis of "the sort of physical conditions and relations that have been assumed to bestow upon an internal state the status of representation" (Ramsey, 2007, p. 189). This focus on the thing instantiating some property, satisfying some definition, or meeting some criteria to count as a representation is the defining characteristic of the standard approach. And this is what my approach will abandon, as

[5]This is, again, in the case of psychological rather than neural representation, but Burge glosses his non-naturalistic definition like so: "Representational psychological states are those that have veridicality conditions as an aspect of their natures — as an aspect of the fundamental explanation-grounding kinds that they instantiate" (Burge, 2010b, p. 2-3). Needless to say, Burge's view is nothing like the Cartesian view I illustrated my point with.

I'll describe in sections 3.4 and 3.5.

The final thing I want to note is this: 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 metaphysics. A quick look at any neuroscience journal will show plenty of concern for representations, and virtually no concern with the kind of debates or objections that a metaphysics of representation would have to tackle, like the question whether one's definition of representation includes things that are (or are arguably) not representations.[6] 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 investigate representational "claims," e.g., claims to the effect that some neural activity represents some stimulus. And they say that their goal is "to give an explanatorily and predictively useful account of what it means to make those claims" (Eliasmith & Anderson, 2003, p. 5). They could do this (as I will) without entering into metaphysical waters, but instead they give the following characterization of their task: "The main problem regarding mental representation, both historically and for contemporary philosophers of mind, is to determine the exact nature of the representation relation; that is, to specify the relation between, and representationally relevant properties of, things 'inside the head' and things 'outside the head"' (Eliasmith & Anderson, 2003, p. 5). Other neuroscientists, who are satisfied with a workaday notion of representation, are not my target here. My targets are the philosophers and neuroscientists who think that a definition or metaphysics of representation is a prerequisite for understanding representational explanation. This is what I mean by the standard approach.

The standard approach has issued in many (many) theories of neural representation. I won't rehearse them here, but these theories tend to come with serious problems, puzzles, and counterexamples. And, more importantly, the definitions they propose have not resulted in fruitful interdisciplinary work. Neuroscientists drawing on the philosophical tradition do not generate answers that philosophers working within the standard approach take seriously, and philosophers taking

[6] In fact, an important benefit of my account, which I'll discuss in section 3.6, is that it brings the philosophical debate about representation into closer contact with the debates neuroscientists actually care about.

the standard approach do not make interventions that neuroscientists feel compelled to heed. The puzzles and problems of philosophers seem only to enforce disciplinary silos. Millikan, e.g., sees this in debates over Swampman (Millikan, 2010), and Burge sees it in debates over the disjunction problem (Burge, 2010b, p. 324).

To be sure, that last paragraph is far from a dispositive argument against the standard approach in any of its many incarnations. I flag these difficulties only as motivation, as reason to be open-minded about non-standard approaches. So I'll set aside the problems for the standard approach and focus instead on the fact that alternatives to it exist. Fleshing out one such alternative is the main purpose of this paper. I'll turn to that task shortly, but first I'll pause to clear up some common misconceptions about my approach.

3.4 Methodological nominalism

My approach will be to forget entirely about the property of representation, and elucidate representational explanation through a discussion of the way *representational notions themselves* figure into the explanatory economy of cognitive science. Call this approach *methodological nominalism*. The task for a methodological nominalist about neural representation is to explain how representational notions support cognitive scientific explanations, without defining or making any commitments about a property of representation.

If the methodological nominalist is successful, she will have answered our questions about representational explanation without detouring through the increasingly baroque debates over the existence and nature of the property of representation. So, in addition to providing an account of representational explanation, she will have cast doubt on approaches that make the resolution of those debates a central task. It is implicit in this that methodological nominalism is not a traditional scientific anti-realism. Traditional scientific anti-realism is a metaphysical view. Methodological nominalism is, of course, a methodological view. 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, is irrelevant to our inquiries.

But even setting aside the methodological/metaphysical distinction, 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*. Even construed metaphysically, methodological nominalism would be about neither: it has no qualms with the existence of 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 the way 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, (Hacking, 1983, p. 23) as opposed to a traditional anti-realism like van Fraassen's (van Fraassen, 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, BEING A POSITRON, exists. And, more to the point, it leaves it open whether our philosophical understanding of physics depends on our defining that property or giving an account of its nature. The methodological nominalist is not even questioning the legitimacy of claims like "the positron left such-and-such a trace" — she is questioning the necessity, for understanding this claim, of investigating the property BEING A POSITRON.

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 are happy for representational theories to be true, because we think their truth 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.

[7]This is in sharp contrast to the way some, like Thomson and Piccinini (2018), frame anti-realism about representations. And even philosophers who don't explicitly frame their realism as entity realism often defend it with arguments for entity realism. E.g., see Ramsey's argument for realism from the fact that representations have causal properties (Ramsey, 2021, p. 62).

If we think about models rather than theories there is even less reason to worry: the accuracy, fruitfulness, explanatory force, etc., of models is not in question. What's under scrutiny is the necessity of understanding that accuracy (etc.) by appealing to a definition of the property of representation. So methodological nominalists can accept, and even aim to support, everything cognitive science uses representational notions to say about the causal structure of the brain. This is not a Dennettian sort of instrumentalism, which allows that the causal structure of the brain may be different than the representational story would have it, and that representational notions are of use only to summarize, systematize, or predict behavior. I certainly won't be arguing that "all there is" to representing is behaving in a way that is predictable from the intentional stance (Dennett, 1989, p. 29). And, more to the point, I won't be arguing that all representational notions do is allow us to predict behavior. They help us to capture the real causal structure of the brain.

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." I'll have more to say about this in the final section — for now, on to the positive account of representational explanation.

3.5 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 the relevant 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 it* to the model to help you impose, on the 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 1 in these conditions" — describing the system not in terms of its own properties, but literally *in terms of* the model you think of it as representing, and whose structure you want it to mirror.[8] Thinking of the circuits as representing parts of the model, and thereby thinking of them in terms of the model, forges an intuitive and useful connection between model and target system.

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 saying 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

[8]To preempt an objection, this need not mean that the computer actually represents those logical operations and the entities they are defined over. This would cause many problems of indeterminacy. It's easy to interpret a logic gate so that it represents its dual, e.g., AND rather than OR (Sprevak, 2010). It's easy to 'carve up' physical states in different ways, and on different carvings the states represent and compute vastly different functions (not just one function and its dual) (Shagrir, 2001). And inputs and outputs would have to be arbitrarily stipulated to represent (say) 1s and 0s, rather than Ts and Fs and so on. These results are not acceptable if you think the gates are really representing something, and that it is the gates' really representing that something that allows you to impose structure on them.

This objection is also irrelevant in an important sense: as I have emphasized, the point I'm making is methodological. The question is whether we need to define representation to understand representational explanation, and this objection doesn't address that question.

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 model and target system 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*), 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 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.

Note 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", an ancient Greek device that calculated astronomical relationships. Since it was discovered, its inputs and outputs, as well as its internal causal structure, have been understood representationally and modeled using structures defined over astronomical entities. E.g., we see debates over models of some gear train in the mechanism — whether to model it with *this* function or *that* one — cast as debates over what the gear train represents — *this* relationship or *that* one. And, in line with the above, this representational thinking licenses descriptions of the mechanism *in terms of* the domain the models are defined over. A function of the mechanism that is modeled by mathematical 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). With regard to internal structure, gear trains are described similarly: "the motion of the sun [is] subtracted from its lunar equivalent" (Marchant, 2008, p. 148). 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 to 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 mathematical 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).

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 notice 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). Those capacities, like in the cases above, are generally understood as abilities to produce certain environmentally-defined outputs or responses to environmentally-defined inputs, stimuli, or states of affairs more broadly.[9]

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

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

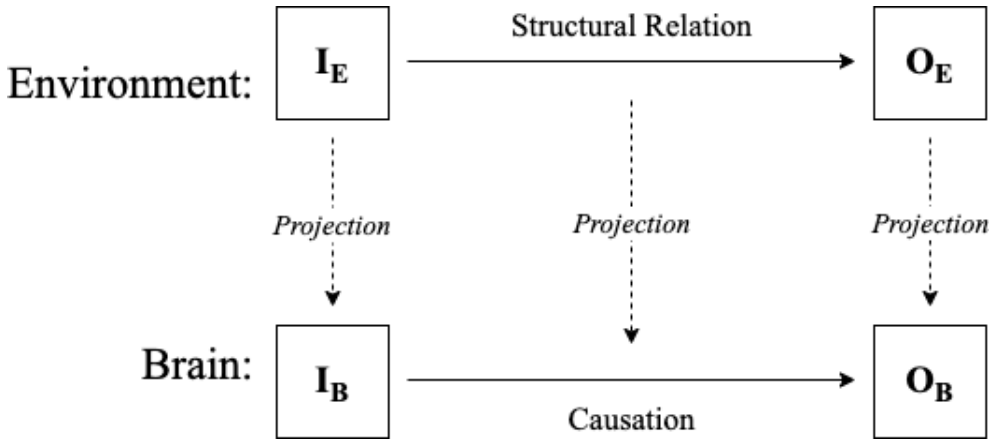

Figure 3.1: A schematic showing the relation between one environmental state (I_E) and another (O_E), and its projection onto the brain's causal structure as a model of the brain's transition from an internal state (I_B) corresponding to the first environmental state to an internal state (O_B) corresponding to the second.

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 actually belonging to the category *face* or *non-face*.[10] 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 the brain to its task domain and to project structures from that domain onto the brain as models of its causal structure (see Figure 3.1).[11] 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 is a simpler case than most, 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

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

[11] Compare Egan (2014) in relation to the figure, and see section 3.6 for a brief discussion of her view.

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 most efficient route back from its final destination, which the animal tends to take. Because it moves from the former to the latter environmentally-defined entities or states — from a set of distances and directions travelled to a new route taken — the animal must move between internally-defined states corresponding to those environmentally-defined ones. It must have a causal structure can be modeled with reasonable accuracy by a function from the distances/directions travelled to the most efficient path home, and by some algorithm or process that computes that function.

Let me make a caveat and a qualification. The caveat is that I intend the notions of computation and algorithm quite broadly, simply defined as any description of a sequence of states for implementing some function.[12] The formal description could be a computer program, a dynamical equation describing continuous change 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, 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 an algorithm whose stages or transitions are *themselves* defined over environmental entities.[13] E.g., we do not just model the brain as moving from states corresponding to low-level visual features

[12]Thanks to Rosa Cao for discussion on this point.

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

to states corresponding to the category of the object it perceives. We model it as doing this via algorithms that are themselves defined over further environmental qualities. 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 something as a face/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. We're using the relationships between the features at various levels 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[14] (whether a mathematical or environmental domain) — e.g., the capacity to recognize faces, or to get from one place to another. Otherwise, environmental structures will not 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 causal structures, 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. 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

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

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.-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 probably find useful.[15]

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 such that, if they did accurately model the brain's causal structure, this 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.

Note that, as I've advertised, this is an account of what representational notions allow us to do, and *not* an account of the property of representation or the representation relation. It would, in fact, be 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, n.d.-b) or stating the constraints on the model. This would make any given brain activity represent a huge variety of things, and we would therefore have made no progress on our more fundamental question of representational *explanation*. With the hippocampus, e.g., we would be left wondering why *a particular one* of its virtually infinite representational contents figured so successfully into our explanation of how the hippocampus supports spatial navigation. See, e.g., Ramsey (2007, pp. 190–203) tussling with this sort of question, in his case a question about whether indicator representations can *really be* representations, since they don't look enough like other representations and since being an indicator isn't distinct enough from other causal roles or similar enough to other ways of being a representation.

[15]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, and they need not be consulted to understand how and why representational explanation works. They only elucidate some conditions under which it is likely to be fruitful.

But instead of investigating the property of representation, I have been describing the way representational *notions* help cognitive scientists in their explanatory and modeling tasks. Representational notions do this not by referring to some property that figures importantly into the explanations or models, but by providing tools for constructing and understanding models. The point is that we've answered the question of how and why representational explanation works without *even having to enter* these debates about what representation is — debates that even sophisticated and like-minded accounts, like Ramsey (2007), find themselves mired in. There is no question whether indicator representations *are really* representations; there is only the question of whether representational notions help us think about and model specific systems along the lines of the account above. If the answer is yes — if representational notions are useful for this purpose — then we should think about and model those systems using representational notions, and we will have a case of representational explanation. If the answer is no, we shouldn't and we won't. Note also that because of all this, my account is only distantly related to isomorphism theories of representation. Aside from the difference noted above — that my account is not an account of the property or relation of representation — there is also the difference that (approximate) isomorphism plays only a minor role on my account, corresponding to a common sense requirement that *models* be approximately isomorphic, in explanatorily relevant ways, to their target systems. What makes a representational explanation appropriate is not, in the main, an isomorphism between two systems. To gloss what I described above, representational explanations will be appropriate when *our goal is to explain a system's relations to its environment or its environmentally-defined behavior, and environmental structures provide models that serve this goal for the system and explananda at hand*.

An important upshot of all this 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. And philosophers and scientists who are interested in representational explanation can focus on the more fruitful and tractable project of understanding representational explanation *as a form of explanation*, rather than as a metaphysical commitment.

This also puts philosophers in a position to join the neuroscientific debates. The focus on representational explanation *qua* mode of explanation connects us directly to neuroscience's own concerns about when and how to use representational explanations, in a way that Swampmen and disjunction problems do not. Though some work in neuroscience appears to be on the metaphysics of representation, it can be interpreted along the lines of my view (Baker, Lange, et al., 2021; Baker, Lansdell, et al., 2021; Poldrack, 2010). But more importantly, the more common debate in neuroscience is not about what representation is. It is about the utility of representational explanation in particular cases. Neuroscientific anti-representationalists do not have problems with the property of representation, but simply prefer other modes of explanation and types of models for the usual scientific reasons. Their anti-representationalism isn't a metaphysical commitment but a policy (cf. Conant, 1952) against a certain type of explanation, for reasons to do with its explanatory fruitfulness. This is, e.g., how Shenoy et. al. (Shenoy et al., 2013), along with the rest of the motor control community, understand the debate between representational and anti-representational approaches to motor cortex. This is a long way from a common sort of anti-representationalism in philosophy, exemplified by Chomsky's eliminativism (Chomsky, 1995) and Hutto & Myin's dynamicism (Hutto & 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. This interdisciplinary bridge is an important benefit of setting aside the metaphysics of representation, allowing the philosophical discussion to track a debate of real significance to cognitive science.

Before moving on, 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 relation. 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, whence their explanatory success?

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 factors 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 function of interest but (in the normative terms that representational thinking allows us to use) 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., when we dismiss crystals as *imperfect* which do not fit the prototypes described by our best mineralogy (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; but most imperfect crystals aren't, for most of our purposes. So misrepresentations, on the account I've given, are not necessarily idealized away. But naming something a misrepresentation is still a way of saying that it is 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.[16] 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 a more illuminating one: 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

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

understand face-discrimination as just a special case of expert discrimination (Kanwisher & 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 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 & McNaughton, 2006). Then general scientific criteria will issue in a straightforward endorsement of modeling the hippocampus as representing (i.e., with structures borrowed from the relations between) an animal's current spatial location and properties of the environment, rather than its intended direction of movement and other properties.

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 using representational notions allows us to do. It allows 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. And the way it does this does not depend on the brain's structures or activity instantiating some property BEING A REPRESENTATION, or BEING A REPRESENTATION OF X. Looking at what representational notions let us do is simply more revealing than looking at the nature of the property they may refer to — in fact, we can learn all we want to know about

representational explanation without ever discussing that property.

3.6 Upshots, objections, and miscellany

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 *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. Representation is not something the brain does, or a privileged relationship neural activity itself has to certain things. Representation is a notion that helps cognitive science model and understand the brain. The important questions are about how it does this, and they can be answered by focusing on the role of representational notions themselves in cognitive science, particularly the role of those notions in modeling the causal structure of the brain. It is no objection to point out that neuroscientists say the brain represents such and such, ask where its representations are, and so on (cf. Bechtel, 2016; Ramsey, 2021).[17] As I said in section 3.4, those are questions the methodological nominalist has no problem with. The problem is with philosophical tendency to go on from here to investigate the property that "representation" refers to in these sentences. But if we start from the more basic questions for the philosophy of cognitive science, about how and why representational explanation works, we bypass that issue altogether. In what remains I will discuss a few objections to this approach, some of its upshots, and its connections to related views.

First, consider a broad objection, similar to one Burge makes of Dennettian instrumentalism:

[17]Though this is a sensible objection to an *entity* anti-realism about representations. To be fair, Bechtel's broader point in that paper is that representational notions play a role in experimentation (especially as targets to be discovered and manipulated) as well as model-building. Of course this is correct, and 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? Read in this light, Bechtel (2016) strongly complements the account I've given here.

"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, 2010a, p. 3). I've distanced myself from instrumentalism, but there is a similar inference to make about neural representation that might challenge my approach: 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. In cognitive science we are confronted, in the first place, with patterns of behavior that we describe and explain in representational terms. The objection says that to understand these descriptions and explanations, we must accept the existence of a property of representation, and commit to defining it. The account in the previous section is a counter-example to precisely this idea — 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 relevance of those properties for understanding cognitive scientific explanation. So if the inference is valid there must be a further reason for it.

Here is one potential reason: representational notions help us build causal models of the brain, but you might think we need to specify something else about our models: what *level* of causal structure they are supposed to capture. Not causal structure at the level of atoms, rarely at the level of synapses, sometimes at the level of neurons, often at the level of neuronal populations, ... How can we specify this level of causal structure, except as *the representational level* — the level at which representations exist? This is an intuitive thought, but it has things exactly backwards. Neuroscience doesn't just seek representations, it seeks models of the brain that explain behavior. Representations would be useless on their own — saying that the brain represents something only forms even the beginning of an explanation if we have at least a sketchy causal model in mind, to the effect that whatever structures implement that representation have an effect on behavior. What we are looking for in the first place are these causal structures, the sources of behavior. Those sources will be found at all kinds of different levels, and there need be nothing they share in advance, except that they will eventually explain behavior. Where it is fruitful to use representational

notions as we construct our models, we will. The system needn't have a *representational level* pre-installed for this strategy to make sense, any more than it must have *calculus levels* pre-installed for calculus to be a useful tool in modeling the brain.

Perhaps the objection could say instead that some interesting patterns are common between different systems, and representational notions tell us what exactly those systems have in common: they have the same representations. That is, to be sure, one way of explaining what is going on when we say things like, "both primates and rodents use egocentric representations in navigation" or, for that matter, "these two rats use the same representations to get through the maze." (These are not sentences you'll see in neuroscience journals — I'm trying to be as generous as possible to the objection.) But consider all the resources the methodological nominalist can use to explain this kind of claim. She can talk about modeling, about the causal vehicles we're thinking about representationally, about the broader causal structures those vehicles are embedded in, and so on. Is there really nothing the methodological nominalist could do with all that material? She might say that these comparisons serve to *correlate* the causal structures of the two systems, to compare models of the two systems by identifying parts of the models that can be overlaid on each other to reveal similar or interestingly different structures. Compare: you're from New York, the home of Central Park. I'm from San Francisco, and I tell you that we have our own version of Central Park, called the Golden Gate Park. What am I doing here, and how should you understand me? Certainly not as invoking a property, BEING A CENTRAL PARK-LIKE PARK, and claiming that it subsumes Golden Gate Park. What I'm doing is identifying a point at which you can overlay the maps of the two cities and see similarities in their structure, or interesting types of dissimilarity — either internal to the park, in the park's relations to and interactions with the rest of the city, or perhaps in other respects (in the history of the parks, e.g., if Frederick Law Olmsted had designed both). If you were trying to understand my comparison and found yourself interrogating me about how to define the property BEING A CENTRAL PARK-LIKE PARK, you would know you had lost the thread of the conversation.[18]

[18]To be fair, this is the kind of thing philosophers tend to do when we get together at the bar. But that doesn't mean it's a good strategy, just that philosophers are bad drinking buddies.

Prima facie, at least, comparisons using representational notions seem to do something similar. I'm telling you something important about the two rats' brains, but it's about similarities (or perhaps interesting differences) in either their internal causal structures or in their relations to the environment and the kind of features in the environment they are responsive to. To be clear, this is a suggestive and partial account at best. Capturing what goes on when we make comparisons using representational notions would take (and would be worth) a longer discussion. But clearly this is a phenomenon to be explored, not an argument against the methodological nominalist.

Another objection might point out apparently harder cases. Linguistic representation comes up often in this context. It seems, prima facie, more difficult to hold that our representations of language, a representational domain itself, don't inherit some of language's representational properties. But on a closer look it's not clear why, or especially why those representational properties would be relevant to our explanations of language use. In fact, even in linguistics a commitment to representation is not typical, either classically (cf. Chomsky, 1995) or in more recent work (cf. Adger, n.d.). When they think about the brain, linguists tend to think of it as instantiating "structures" (Adger, 2003) without intentional or representational properties. And on a close inspection, language use is actually a parade case for the methodological nominalist. Here we have a task environment of visual and auditory signs, in which we find incredibly rich, detailed, and well-investigated structures — natural language grammars. And we are interested in how the brain gets around in those structures, particularly since it likely hasn't had the evolutionary time to come up with any dynamical short-cuts. That is, we have a case where we are interested in explaining how the brain supports environmentally-defined capacities in an environment with rich and accessible structure, and we have every reason to believe that structure will model, in interesting and informative ways, the brain's causal structure. As I discussed in section 3.5, the methodological nominalist would predict that representational notions will be a particularly useful resource in a case like this.

Maybe what's intuitively more difficult about language is that intentional notions come into the explanandum. What's up for explanation is *understanding* itself. There is a Searlean view in the vicinity of this idea, but a more current version might be that the behavior we're trying to model

isn't explained just by showing how it came about. We also have to show how it was counterfactually dependent on what it represented. So the models should be counterfactually robust. That much is fine for the methodological nominalist, because it's a fact about causal models and what we want them to explain, what causal structures would be explanatory, and so on. But perhaps the behavior is also counterfactually dependent on other states *representationally-described*? So, e.g., we might say you saw something *as* a snake. Your future behavior will depend on this intentional fact — if you had seen it as a rope you wouldn't have run away from it or screamed. But this, too, is explicable in terms of the way representational notions characterize causal structures. If we're interested in the way subjects perceive and respond to snakes as opposed to ropes, we will show them some ropes designed to test which features drive different types of response: jumping and shouting, on the one hand, and disinterest or tying a clove hitch on the other. In cognitive science, to say that a subject sees a rope as a snake is not to claim that the correct categorization of her state is as a REPRESENTATION OF A SNAKE in some metaphysically robust sense, but a way to flag that salient parts of the causal structure responsible for the subject's response to the rope are the ones more typically involved in responses to snakes. This kind of comparison can be interesting and important, and it can shed light on innate object-detection mechanisms and their relation to the environment or to evolutionary history (Lobue & Deloache, 2011). But it does not pose a problem for the methodological nominalist, who has the tools to explain its import and explanatory role.[19]

Are there other cases, likely to be more difficult for the methodological nominalist? I've looked primarily at complex cases where representational notions don't just characterize some causal structure but are central to the discovery and investigation of that structure, not to mention experimentation on it. I've looked at cases where computations, like the path-integration algorithm, are defined over the states we characterize as representations. I've dealt with cases where both

[19]Another common objection is that representationally-described states are *explananda*, and so representational description can't be characterized just as an explanatory tool. But the way representational notions characterize explananda is not terribly mysterious. As with the case above, they serve to categorize patterns of behavior into groups that we expect to have the same or interestingly similar causal sources. If we describe our explananda as a subject's *seeing the rope as a snake*, we are not saying that our explananda is this taxonomic fact: the subject's state falling under the category, SEEING X AS A SNAKE. We are merely stating in advance that the subject's behavior is of a kind that is usually brought about by the causal structures that mediate responses to snakes, and that we expect a good explanation to model the current phenomenon as being brought about by the same or similar structures.

behavior and brain activity have significant counterfactual dependence on the parts of the environment we understand them in terms of, and we've seen that the account applies just as well to cases where there is counterfactual dependence on other structures *representationally-described*. I stuck to cases where the use of representational notions is deeply ingrained, highly informative (causally, counterfactually, and in other ways), and seemingly intractable. If there is some other feature of a case that would make it particularly difficult for the methodological nominalist, I haven't been able to divine it. That's not to say it doesn't exist. But while I wait for harder cases to be proposed, I'll move on.

So far I've talked about ways of fleshing out the objection I loosely connected to Burge. I've focused on specific features of cognitive scientific explanation that might justify an inference to the reality and relevance of the property of representation. So far as I can see, there is little to be said for this type of argument. So let me step back and discuss two broader concerns. The first involves an indispensability argument. I haven't said that representational notions are indispensable in cognitive science, but I've certainly suggested they're difficult to dispense with. It is tempting to move straight from here to the standard approach. But I think this move turns on a mistake that is familiar from section 3.4: confusing the debate about the property of representation with a debate about entity-realism. If explaining the brain means talking about some particular *stuff*, anti-realism about that stuff looks dubious. But if we have to talk about it *in certain terms* or *using certain notions*, that might be simply because those terms and notions are the most useful given our explananda (along with our psychology, our pragmatic concerns, and so on). This is going to be common in modeling practices more generally. We model a neuron using calculus not because it satisfies some definition of what it takes for a bit of physical stuff to perform calculus operations, but because calculus helps us latch onto causal structures that we find to be explanatory in that particular case. The choice of models, and especially of the terms they are couched in, is rarely a matter of metaphysics.

The second 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 a *person* in the sense that *person* is a useful notion with which to conceive of 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 say things in physics like, "The universe is made up of particles/fields/…," and understand those as important categories needing definition? Science would be effectively barred from taxonomy, except taxonomy understood in highly pragmatic terms.

In both the scientific and non-scientific cases, I think the answer is contained in the questions: "Wouldn't it be troubling if …"; "Wouldn't it be strange if …". Good: so say what's troubling about it. If methodological nominalism about personhood is troubling, that is presumably because a theory of personhood-ascription that neglected actual personhood would fail to meet some desiderata. That means the methodologically nominalist approach to personhood would fail to meet some desiderata, and could be rejected on those grounds. The point is to say what desiderata it fails to meet. In the case of personhood it seems to fall short at least of our desire to ground moral status — you don't have moral status because I think of you a certain way, but because you are a certain something. So, insofar as we want our account of personhood to ground moral status, we have an argument against methodological nominalism in that case.

What about scientific categories, then? I think methodological nominalism is a salutary approach in many cases. But where it isn't, it will be because there is some desideratum that it doesn't meet. It may be that we understand physics as having the ultimate goal not of modeling the way the universe brings about certain states of affairs, but of taxonomizing its basic constituents, and perhaps saying how they relate to each other and compose other states of affairs. Likewise, perhaps, with parts of chemistry and materials science. I don't want to commit to this understanding of physics or chemistry. 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 will be fairly common.[20]

[20] And note again that it doesn't keep us from talking about representations, protons, or whatever else. The methodological nominalist has no problem saying that these things exist and giving explanations in terms of them. The problem comes only when we try to understand these claims or explanations by appeal to a metaphysics of the properties they instantiate.

I have one more issue to discuss before concluding. Everything I've said in this section is complicated by another view, very much in the vicinity of mine, called pragmatism or deflationism (R. Cao, 2022; Egan, 2019, 2021; Mollo, 2020). Deflationists accept the standard approach, and offer a view of neural representation: they answer the question of what neural representation is with a deflationary definition or metaphysics. A deflationary metaphysics is distinctive in its sparseness and interest-relativity, issuing in a set of answers similar to those I gave in the first paragraph of this section. The deflationist and I would agree, e.g., that debates over whether the Watt governor really represents are exactly as pointless as they seem. But deflationism nonetheless addresses itself to the question, *what is neural representation?*

Look at the similarities another way: a deflationist could adopt my account, and simply add what I've called elsewhere a "metaphysical appendix" (Richmond, n.d.-b): a definition to the effect that whatever is treated representationally in the way I've described *just is* a representation. They would end up with a deflationary, pragmatic, or otherwise "light" (Egan, in conversation) metaphysics of representation. This would allow them to hold on to the standard approach, and to answer the objections above along the lines of the standard approach — e.g., the "level at which there are neural representations" *was* there all along, because it's just the level at which we were going to end up applying representational notions. The metaphysical appendix will let you do this straightforwardly, and seemingly 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 would be mistaken. I've given you the account post-appendectomy, and I think there is nothing to gain, and much to lose, from opening it up to shove in an organ.[21]

First, by setting aside metaphysics altogether we put our focus on the actual features that make representational explanations successful. Even if the metaphysics of representation can be given a deflationary treatment, that metaphysics is not where the action is. So focusing on the metaphysics makes one liable to miss the important features of representational explanations. Egan, e.g., treats them as glosses on causal models (Egan, 2018). There are important advantages to this

[21] Mollo (2021) is one deflationist who, I think, agrees with me in this respect. Though he takes a quite different tack than me, I understand him as rejecting the question of what it is for some structure or activity in the brain to represent.

view over the usual views of representation, but as an answer to the question of *what neural representation is* it's an unsatisfying place to stop. Though it may answer that question, it makes only limited progress on the more basic questions that motivate our inquiry in the first place, about how and why representational explanation works. On a methodologically nominalist view, however, Egan's view is not an unsatisfying conclusion but a novel and promising *beginning* to a deeper investigation of how the notion of representation serves cognitive science. And, as I mentioned in the previous section, that investigation is one that philosophy shares with cognitive science — an important interdisciplinary bridge that methodological nominalism, and my account specifically, makes available.

Let me conclude with two dialectical reasons to take the appendectomy. First, explicitly eschewing metaphysics makes it less tempting for opponents to complain, as many do about deflationism, that its metaphysics allows too many things to be representations, or makes something's status as a representation problematically subjective. With the property of representation out of view, this type of objection can easily be seen to miss the point, and to re-focus without argument on the irrelevant question of the metaphysics of representation. A number of other worries about deflationism can likewise be dismissed on the methodologically nominalist approach, especially those that try to show that deflationism is a form of anti-representationalism (Hutto & Myin, 2021) or anti-realism (Neander, 2015). I think the deflationist should respond by simply coming over to the methodologically nominalist's way of seeing the debate. For the methodological nominalist, representationalism isn't a commitment to the reality of some property of representation. It's a commitment to representational *explanation*: an explanatory strategy supported by good, pragmatic, scientific reasons. If the deflationist makes this move, the criticisms of Hutto and Myin (2021) and Neander (2015) simply miss the point. We're representationalists, not anti-representationalists. And we make don't challenge — or even *address* — the reality of representations, and certainly not the reality of the *property* of representation. We simply think that, for someone interested in understanding representationalism and representational explanation, understanding that property is not an important task.

And second, this account and it's want of metaphysical commitments raise an obvious challenge to the deflationist *and* her opponents, insofar as they all subscribe to the standard approach. Proponents of that approach must justify it in light of its lack of independent light to shed on representational explanation and the many challenges, puzzles, and distractions from other interdisciplinary work that it entails. They must find some legitimate question in the philosophy of cognitive science that we stand to answer through a metaphysics of representation (deflationist or otherwise) and that cannot be answered without one. Otherwise, we have a good account of representational explanation that does not appeal to the property of neural representation, so what would a theory of that property be for?

Coda

Like I said: not a terribly long journey. But hopefully worthwhile. To conclude, I want to draw out a couple directions of research that this work opens up. The idea is not to give rigorous arguments for these two ideas but to make them clear, and especially to make their implications stand out. So I'll follow the example of my preface, and be a bit more doctrinaire and optimistic than I probably have a right to be. First, I'll sketch an unexplored empirical dimension of Chapters 1 and 3. And second, I'll discuss an upshot of these views for the philosophy of mind. It will become clear that these are neither of them fully-formed ideas. To be frank, they are primarily here because I want to work on them in the next couple years, and I'm unlikely, in the near future, to have another audience as perfectly captive as you all are. But I do think that, in broad strokes at least, these ideas are natural consequences of the papers you've just read. So they should reflect on the dissertation itself, for better or worse.

First, the empirical dimension. I've argued that philosophers of cognitive science, insofar as they are concerned with the way cognitive scientific explanations work, should focus on the kind of thinking involved in those explanations and what it allows scientists to do, rather than any metaphysical underpinnings for the explanations. Prima facie, at least, this project is more empirically tractable than the metaphysics of computation or representation. To illustrate, consider a parallel to some existing work on *teleological functions*. Cognitive science and biology, not to mention common sense itself, are chock full of teleological explanations — explanations of something in terms of the purpose it serves. Philosophers, of course, see a gap in the literature here concerning what exactly a purpose or function is. Psychologists, on the other hand, are interested in understanding

how these explanations work: what cognitive role do these explanations play for their givers and receivers?

The psychological project is most thoroughly worked out by Tania Lombrozo, who studies the role teleological notions play in common-sense explanations. Some of her work just asks which groups of people prefer teleological explanations as opposed to mechanistic ones (Lombrozo et al., 2007), but she also investigates the kind of categorizations (Lombrozo, 2009) and generalizations (Lombrozo & Gwynne, 2014) teleological explanations support, the way they interact with other kinds of explanation like causal explanation (Lombrozo & Carey, 2006), and the kind of cognitive and practical goals teleological explanations serve (Lombrozo & Carey, 2006).

This is precisely the kind of approach that my discussion of computation and representation motivates. If we want to understand how representational explanation works in cognitive science, rather than asking what counts as a representation or how to characterize that property, we might collect a group of cognitive scientists and assign them to one of two conditions, following Lombrozo's paradigm. One group is prompted with (or prompt to give) explanations of some system in representational terms. The other is prompted with (or prompted to give) explanations of that same system in purely mechanistic terms. Then they are asked questions about the system, about its environment, about its behavior under various circumstances, about its causal structure at different levels of organization and with respect to different tasks, and so on. This is roughly the structure of the experiments in Lombrozo et al. (2007). And this work would have exactly the same goal as Chapter 3, but would use experimental tools rather than observation, reflection, and the analysis of case studies.

What's especially important is that something like this project is getting off the ground in philosophy, but with a catastrophically mistaken starting point. Favela and Machery (n.d.) presented cognitive scientists (and philosophers) with a few made-up studies showing that some neural activity was correlated with some stimulus, or was caused by that stimulus, or etc., and asked those cognitive scientists questions about the systems. But, unfortunately, the questions they asked were exclusively categorical: does the neural activity *represent* the stimulus, does it *carry information*

about the stimulus, is it *about* the stimulus, etc. They found a great deal of disagreement about which representation-related categories any given neural activity fell under, and so they argue that the concept of representation should be "reformed or eliminated from use" (Favela & Machery, n.d., p. 5). The main problem with the experiment is clear: it assumes that what representational notions do for cognitive science is to pick out a property that some things instantiate and others don't. On that assumption, too much disagreement over what counts as having that property could put the notion of representation in serious trouble: it would indicate that there isn't a coherent property we're picking out with representational notions.[1] If there is nothing to do the work that the property of representation is supposed to do, then the concept of representation isn't fulfilling its purpose in picking out such a property. But of course concepts do much more for cognitive science than categorize, and there is no reason that inconsistent categorizations of the kind Favela and Machery find would interfere with representational notions' ability to do the kind of thing I discussed in Chapter 3.

A better experiment would look to other potential functions of representational notions, taking into account how they help build models and frame questions, and what they enable cognitive scientists to do more generally. This kind of work might draw on Lombrozo's paradigm as I described it above, or it might hearken back to some early work in what might now be called experimental philosophy, namely Daniela Bailer-Jones' classic work surveying scientists about their understanding of scientific models. Bailer-Jones didn't just ask scientists to categorize examples as MODEL or NOT-MODEL, but asked them about the characteristics and flaws of models, how they used models and what purposes models are supposed to serve, how modeling figured into their research and thinking more generally, and so on. I'll leave it as an exercise to the reader to review her book, *Scientific Models in Philosophy of Science* (Bailer-Jones, 2009), and imagine how misleading it would be, and how much insight would be lost, if she had instead taken Favela and Machery's impoverished approach. All this to say: one upshot of this dissertation is that there is room for informative experimental work on traditional philosophical questions about representational ex-

[1] Or, at least, we are picking out a menagerie of different properties. For the sake of discussion, we can assume that's just as bad.

planation, but current work in this direction takes an approach that is simply inappropriate to its subject matter.

The second thing I wanted to discuss here is the way my views in the philosophy of cognitive science interact with the philosophy of mind more broadly. Consider the following three assumptions, which characterize much of contemporary philosophy of mind:

> The philosophy of mind should be *naturalistic*.
>
> An essential task in philosophy is to give a *metaphysics of mind*.
>
> The most promising approach in the philosophy of mind is *computational* and *representational*.

I think Chapters 1 and 3 furnish an argument that the three assumptions are incompatible.

First, let me define the three assumptions more carefully. Naturalism comes in different forms, but it generally involves a pro-scientific attitude on the part of philosophers (Bryant, 2020). A weak form of naturalism would say that the philosophy of mind should be consistent with scientific results. But most naturalists have something more in mind: they think philosophical theories should be grounded in or supported by scientific work, or perhaps that they should support that work in some way (e.g., by providing it with foundations). A commitment to souls, ghouls, or reincarnation could be consistent with scientific work, but it would not be naturalistic because it would be too far removed from science to enter into relations of support with it. More substantial notions of naturalism are available, but this minimal one is all I'll need to draw out the problem.

In the philosophy of mind the relevant science is cognitive science, especially psychology and neuroscience. So the weak form of naturalism that philosophers of mind generally accept is the idea that views in the philosophy of mind should do one or both of the following. They should draw support from neuroscientific and psychological work — as, e.g., the isomorphism theory of mental representation might draw support from neuroscientific results indicating that navigation uses map-like representations (Moser et al., 2008). Or they should support psychological or neuroscientific work in some sense — as, e.g., the teleosemanticist about mental representation might think

her work provides conceptual foundations for neuroscientists' reference to neural representations (Neander, 2017).

Moving on from naturalism, the metaphysics of mind is a much older project, but I think it can be summarized straightforwardly. Descartes did not just ask how we could *best frame* theories about the mind. Armstrong did not just argue that causal notions help us *understand* mental capacities. Searle did not just claim that biology helped to *explain* mental capacities, or that computational notions were insufficient to *illuminate* mental processes. In each case, the philosophers were interested in understanding what certain mental capacities *are*, or, more generally, what the mind *is*. Descartes inquired about the "nature and essence" of mind (Descartes, 1984, pp. 210–211). Armstrong argued that mental processes are, or are "identical with," causal processes (Armstrong, 1980, p. 4). Searle argued that cognition cannot *be* computation — he was concerned with the apparent claim of strong AI that "the appropriately programmed computer really is a mind" (Searle, 1980, p. 417). He wanted to know "what is specifically mental about the mental", and to find criteria that ground a "mental-nonmental distinction" (Searle, 1980, p. 420). It was apparently a failure of Strong AI that it did not provide these — it did not describe "the essence of the mental," the qualities in virtue of which something has a mind (Searle, 1980, p. 422). Philosophers like this are after a metaphysics of mind. They may want to know the essence or nature of mind, the necessary and sufficient conditions for being minded, the definition of 'mind,' or what constitutes mindedness. Regardless of their terminology, they want to know what makes minds *minds* — what it is in virtue of which things belong to the category MIND. Or, in the less ambitious but more common cases, they want to know what makes a certain mental capacity the mental capacity it is, or what it is in virtue of which things belong to sub-categories of the mental like PERCEPTION, PAIN, FIRST-PERSONAL THOUGHT, or, for that matter, MENTAL REPRESENTATION. (I'll focus on the more ambitious case of the metaphysics of mind itself to draw out the issue.)

I won't explain the computational/representational approach (CRA) in detail, since I've spent a full paper each on representational and computational approaches. These approaches apply computational and representational concepts, tools, and thinking to a target domain. What's interesting

is what happens when the CRA is combined with the metaphysics of mind. This results in computational and representational answers to metaphysical questions. Focusing just on computation, we get the following:

> The mind is literally a classical computational system — an interpretable, formal, symbol manipulator — of some sort; and cognitive processes, such as reasoning and visual perception, just are classical computational processes of some sort. So construed, [computationalism] is a kind of *empirically motivated, metaphysical* doctrine, in that *it provides a general characterization of what it is to be a mind, or cognitive process*. (Samuels, 2019, p. 106)

> [Computationalism] is not intended metaphorically. [Computationalism] does not simply hold that the mind is like a computing system. [Computationalism] holds that the mind *literally is* a computing system. (Rescorla, 2017, p. 9)

Another good example is Block, who frames the computational theory of mind explicitly as an answer to the question "What is intelligence," the way that H_2O is an answer to the question "What is water" (Block, 1998, pp. 384–385).

So how does naturalism fit into all this? For these views to be naturalistic, on my definition, they must support or be supported by cognitive science. Take the second direction of support first. What would it take for cognitive science to support one of these views? Cognitive science would have to support the claim that the mind is a computing system. If I'm right in Chapter 1, it doesn't do this directly, because while it models the brain with computational tools and resources, it doesn't make any related metaphysical commitments — any commitments to its target systems belonging in the category COMPUTER. Could it support the metaphysical view *indirectly*, then? It's hard to see how. If you need a metaphysics of mind for some reason, is it best to adopt one that adds metaphysical commitments to cognitive scientific language, rather than one that deviates from that language? It is hard to see a case for this. It cannot be for the sake of connecting philosophy and science in relations of support in either direction, because the connection would be illusory —

the words are the same but the substance is not. We can borrow the language of cognitive science if we, too, want to think of the brain or mind in computational terms. But there is no alchemy that turns this into a metaphysical commitment if it isn't already one in cognitive science.

The other option, the other possible direction of support, was for a computational and representational metaphysics of mind to support cognitive science. What would that look like? Maybe the metaphysics of mind provides a *foundation* for computational/representational cognitive science? The problem is that this is precisely what Chapters 1 and 3 were doing without metaphysics. Their goal was to explain how and why computational and representational explanations work. The point was they they work by bringing a particular set of resources and tools to cognitive science, tools that are well-suited to cognitive science's explanatory goals. What more support can a metaphysics provide? What grounding is needed, if we already understand how and why computational explanation works?

Perhaps it is not *how* computational explanation works, but the peculiar *success* of cognitive science that calls for explanation, which explanation must be given in metaphysical terms? This too is dubious. If virology provides tools for thinking about disinformation, and this leads to successful social science, that doesn't support the philosophical claim that disinformation *is a virus*. It would certainly license us to *talk* that way; the point is that it wouldn't support a philosophical project building a metaphysics of viruses to encompass disinformation. Of course, you could imagine a "no miracles" argument to the effect that, if virological notions are so useful in the study of disinformation, there must be some reason. Compare, e.g., Shea's suggestion that a "natural property cluster" must "underpin[] the explanatory purchase of representational explanation" (Shea, 2018, pp. 48–49). But the metaphysical approach isn't the only one that dispels miracles. To reiterate a much-belabored point, it is no miracle that you can use the claw of a hammer to turn a slotted screw, or that you can use a screwdriver to open a can of paint. Concepts, formalisms, and frameworks can be multi-purpose or repurposed just as well as tools can. And just like tools, they will be successful with respect to some purpose when their characteristics are a good match for the relevant goals and subject matter. To understand how the claw of a hammer turns a slotted

screw, we don't need a metaphysics that counts the hammer as a screwdriver or the screw as a nail. To dispel any supposed miracles we need only understand the characteristics of our tools and resources, the characteristics of our materials and subject matter, and the reasons that the former are an appropriate way to achieve our goals with respect to the latter.

Two caveats are necessary here. First, there is more to say about the ways that philosophers might think they are supporting cognitive science. This really calls for a paper-length treatment, but all I'm doing here is fleshing it out as a potential implication of the dissertation. Second, in sketching this argument I've relied on a strong interpretation of my points in Chapters 1 and 3. There, I examined mostly neuroscience. There is more work to do extending those accounts to the rest of cognitive science, particularly to psychology, which might have interesting differences from neuroscience. But I don't see any reason to think that psychology uses representational and computational notions in a more metaphysically-committal way than neuroscience — especially not cognitive neuroscience and neuropsychology, which take up the same explananda as psychology and made up the bulk of my discussion.

The point is that if we grant all this, we sever the naturalistic tie between cognitive science and the philosophy of mind. A naturalistic computational/representational metaphysics of mind is impossible, because its metaphysical aspirations can't be connected to the work of cognitive science in the way that naturalism requires. That would leave a few ways forward for the philosophy of mind, corresponding to which assumption we drop. One, I think, is a dead end. Dropping the CRA but still hoping one's metaphysics is naturalistic means hoping that cognitive science is wrong to use a representational and computational framework. And that might be right, but, as far as I can tell, other approaches in cognitive science, like dynamical systems theory, aren't any more metaphysically-committal, and don't seem to require any more metaphysical underpinning, than computational and representational approaches.

What about the other options? We could give up on metaphysics and understand the philosophy of mind as trying simply to *explain* mental phenomena, rather than validating their subsumption under some category or delineating the criteria for *being* that kind of phenomenon. That, I think,

is what a good deal of exciting work in the philosophy of mind is up to. E.g., to pick just one example close to hand, consider Laurie Paul's work illuminating important features of decision-making, but (as far as I can tell) without making any attempt to set out and defend the criteria for something to be a decision, a decision of a certain type, or so on (Paul, 2014). Or we could give up on naturalism. This would mean that we keep trying to give a metaphysics of mind (doing it in computational/representational terms or otherwise), but we take our constraints from outside of cognitive science, e.g., from phenomenological, sociological, ethical, or normative considerations. Leaving aside the ambitious project of saying what it is *to be minded*, there is important and fruitful non-naturalistic work on narrower questions in the metaphysics of mind, e.g., on what it is to perform or to be capable of perceptual demonstratives (Dickie, 2015). If I'm right about this work, it explicitly targets constitutive questions about categories like thought, mind, reference, etc. — questions about *what it is to belong to one of those categories* — but without the assumption that its answers should have any special relation to cognitive science. There is also plenty of work in the philosophy of mind that doesn't fall straightforwardly into these categories. It is, at least, not immediately obvious whether certain philosophical work on the perception of color (e.g., Morrison, 2020) would count as metaphysical and/or naturalistic in the sense in which those are in tension. It's also possible that work in this area, which appeals to more local and nuanced aspects of cognitive science than just the use of representational/computational explanation, has a better argument that cognitive science really does support certain metaphysical views.

So I'm not arguing for a complete overhaul of the philosophy of mind. There is plenty of work that's unaffected by the criticism I'm making. But I suspect that if I'm right about computational and representational explanation, there is a strong case against any metaphysics of mind that aspires to be naturalistic (especially if it is computational and representational), and generally against the natural way of thinking about the philosophy of mind as a metaphysical and naturalistic project. As I promised, this is a bit sketchy, but I think it's an implication of my treatment of cognitive science. Philosophers using cognitive science for philosophical projects need cognitive science to be saying things that could, at least in principle, support those projects. But philosophers often

make fairly uncritical use of cognitive science, and to see how philosophers can *legitimately* use it we need a more rigorous understanding of cognitive science *as a science*. That more rigorous understanding may undermine certain approaches to the philosophy of mind, but I think it points us in the direction of more promising ones.

References

Adger, D. (n.d.). What are Linguistic Representations? Forthcoming.

Adger, D. (2003). Core Syntax. Oxford University Press.

Allen, C. (2017). On (not) defining cognition. Synthese, 194(11), 4233–4249.

Anderson, M. L. (2014). After Phrenology. MIT Press.

Armstrong, D. (1980). The Causal Theory of Mind. In D. Armstrong (Ed.), The nature of mind (pp. 16–31). University of Queensland Press.

Bailer-Jones, D. M. (2002). Scientists' thoughts on scientific models. Perspectives on Science, 10(3), 275–301.

Bailer-Jones, D. M. (2009). Scientific Models in Philosophy of Science. University of Pittsburgh Press.

Baker, B., Lange, R., Achille, A., Cao, R., Kriegeskorte, N., Schwartz, O., & Pitkow, X. (2021). Generative Adversarial Collaborations Proposal (tech. rep.).

Baker, B., Lansdell, B., & Kording, K. (2021). A Philosophical Understanding of Representation for Neuroscience.

Barack, D., & Krakauer, J. W. (2021). Two Views on the Cognitive Brain. Nature Reviews Neuroscience, 22, 359–371.

Bechtel, W. (2016). Investigating neural representations: the tale of place cells. Synthese, 193(5), 1287–1321.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. Neuron, 100(2), 490–509.

Block, N. (1998). The Mind as the Software of the Brain. In E. E. Smith & D. N. Osherson (Eds.), Thinking: An invitation to cognitive science (2nd ed., pp. 377–425). MIT Press.

Blohm, G., Kording, K. P., & Schrater, P. R. (2020). A how-to-model guide for neuroscience. eNeuro, 7(1), 1–12.

Boden, M. (2006). Mind as Machine: A History of Cognitive Science. Oxford University Press.

Bontly, T. (1998). Individualism and the Nature of Syntactic States. The British Journal for the Philosophy of Science, 49(4), 557–574.

Borghesani, V., & Piazza, M. (2017). The neuro-cognitive representations of symbols: the case of concrete words. Neuropsychologia, 105(June), 4–17.

Brette, R. (2015). Philosophy of the spike: Rate-based vs. Spike-based theories of the brain. Frontiers in Systems Neuroscience, 9(November), 1–14.

Brette, R. (2022). Brains as Computers: Metaphor, Analogy, Theory or Fact? Frontiers in Ecology and Evolution, 10(April), 1–5.

Brillinger, D. R. (2014). ". . . how wonderful the field of statistics is. . . ". In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.), Reminiscences of the columbia university department of mathematical statistics in the late 1940s (pp. 41–47). Chapman; Hall/CRC.

Bryant, A. (2020). Naturalisms. Think, 19(56), 35–50.

Burge, T. (2010a). Origins of Objectivity. Oxford University Press.

Burge, T. (2010b). Origins of perception. Disputatio, 4(1), 1–38.

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

Butler, K. (1998). Content, Computation, and Individuation. Synthese, 114(2), 277–292.

Cao, R. (2019). Computational Explanations and Neural Coding. In M. Sprevak & M. Columbo (Eds.), The routledge handbook of the computational mind (pp. 283–296). Routledge.

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

Carandini, M. (2012). From circuits to behavior: a bridge too far? Nature Neuroscience, 15(4), 507–509.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13(1), 51–62.

Chalmers, D. J. (1996). Does a Rock Implement Every Finite-State Automaton? Synthese, 108, 309–333.

Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition. Journal of Cognitive Science, 12, 323–357.

Chang, L., Breuninger, T., & Euler, T. (2013). Chromatic Coding from Cone-type Unselective Circuits in the Mouse Retina. Neuron, 77(3), 559–571.

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. Cell, 169(6), 1013–1028.

Chirimuuta, M. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. Synthese, 191(2), 127–153.

Chirimuuta, M. (2019). Charting the Heraclitean Brain: Perspectivism and Simplification in Models of the Motor Cortex. In M. Massimi & C. D. McCoy (Eds.), Charting the heraclitean brain (pp. 141–159). Routledge.

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

Chomsky, N. (2000). Explaining Language Use. New horizons in the study of language and mind (pp. 19–45). Cambridge University Press.

Churchland, P. S., & Grush, R. (1999). Computation and the Brain. In R. Wilson & F. C. Keil (Eds.), The mit encyclopedia of the cognitive sciences (pp. 155–157). MIT Press.

Cisek, P. (1999). Beyond the Computer Metaphor: Behaviour as Interaction. Journal of consciousness studies, 6(11), 125–142.

Cobb, M. (2020). The Idea of the Brain. Basic Books.

Conant, J. B. (1952). Modern Science and Modern Man. Doubleday.

Conway, B. R., Chatterjee, S., Field, G. D., Horwitz, G. D., Johnson, E. N., Koida, K., & Mancuso, K. (2010). Advances in Color Science: From Retina to Behavior. Journal of Neuroscience, 30(45), 14955–14963.

Conway, B. R., Eskew, R. T., Martin, P. R., & Stockman, A. (2018). A tour of contemporary color vision research. Vision Research, 151(August), 2–6.

Cornelissen, F. W., & Brenner, E. (2015). Is adding a new class of cones to the retina sufficient to cure color-blindness? Journal of Vision, 15(13), 1–7.

Cummins, R. (1983). The Nature of Psychological Explanation. MIT Press.

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

Cushman, F. (2020). Is Cognitive Neuroscience an Oxymoron? In A. J. Lerner, S. Cullen, & S.-J. Leslie (Eds.), Current controversies in philosophy of cognitive science (pp. 121–133). Routledge.

Dacey, D. M., & Lee, B. B. (1994). The 'blue-on' opponent pathway in primate retina originates from a distinct bistratified ganglion cell type. Nature, 367, 731–735.

Daniel, R., Schuck, N. W., & Niv, Y. (2015). How to divide and conquer the world, one step at a time. Proceedings of the National Academy of Sciences of the United States of America, 112(10), 2929–2930.

Danks, D. (2019). Probabilistic Models. In M. Sprevak & M. Colombo (Eds.), The routledge handbook of the computational mind (pp. 149–158). Routledge.

Dennett, D. C. (1989). The Intentional Stance. MIT Press.

Dennett, D. C. (1994). Cognitive Science as Reverse Engineering: Several Meanings of "Top Down" and "Bottom Up". In D. Prawitz, B. Skyrms, & D. Westerstahl (Eds.), Logic, methodology and philosophy of science ix (pp. 690–689). Elsevier Science.

Derrington, A., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus. Journal of Physiology, 357, 241–265.

Descartes, R. (1984). Principles of Philosophy. In J. Cottingham, R. Stoothoff, & D. Murdoch (Eds.), The philosophical writings of descartes, volume 1 (pp. 193–292). University of Cambridge Press.

Devalois, R., & Devalois, K. (1993). A Multi-Stage Color Model. Vision Research, 33(8), 1053–1065.

Devalois, R., & Devalois, K. (1997). Neural Coding of Color. In A. Byrne & D. R. Hilbert (Eds.), Readings on color volume 2 (pp. 93–140). MIT Press.

Dewhurst, J. (2018). Individuation without Representation. The British Journal for the Philosophy of Science, 69, 103–116.

Diamond, C. (1978). Eating Meat and Eating People. Philosophy, 53(206), 465–479.

Dickie, I. (2015). Fixing Reference. Oxford University Press.

Dulai, K. S., Von Dornum, M., Mollon, J. D., & Hunt, D. M. (1999). The evolution of trichromatic color vision by opsin gene duplication in new world and old world primates. Genome Research, 9(7), 629–638.

Dunbar, K. N. (2002). Understanding the role of cognition in science: the Science as Category framework. In P. Carruthers, S. Stich, & M. Siegal (Eds.), The cognitive basis of science (pp. 154–171). Cambridge University Press.

Edsel, A. (2016). Breaking Failure. FT Press.

Egan, F. (1991). Must Psychology Be Individualistic. Philosophical Review, 100(2), 179–203.

Egan, F. (1992). Individualism, Computation, and Perceptual Content. Mind, 101(403), 443–459.

Egan, F. (1994). Individualism and Vision Theory. Analysis, 54(4), 258–264.

Egan, F. (1995). Computation and Content. Philosophical Review, 104(2), 181–203.

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

Egan, F. (2003). Naturalistic Inquiry: Where does Mental Representation Fit in? In L. M. Antony & N. Hornstein (Eds.), Chomsky and his critics (pp. 89–104). Blackwell Publishing.

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

Egan, F. (2012). Metaphysics and Computational Cognitive Science: Let's Not Let the Tail Wag the Dog. Journal of Cognitive Science, 13(1), 39–49.

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, & T. Schlicht (Eds.), What are mental representations? (forthcoming). Oxford University Press.

Egan, F. (2019). The nature and function of content in computational models. In M. Sprevak & M. Colombo (Eds.), The routledge handbook of the computational mind (pp. 247–258). Routledge.

Egan, F. (2021). A Deflationary Account of Mental Representation. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), What are mental representations? Oxford University Press.

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

Euston, D. R., & 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., & Machery, E. (n.d.). The Untenable Status Quo: The Concept of Representation in the Neural and Psychological Sciences. Draft.

Feldmann, M., Beckmann, D., Eysel, U. T., & Manahan-vaughan, D. (2018). Early Loss of Vision Results in Extensive Reorganization of Plasticity-Related Receptors and Alterations in Hippocampal Function That Extend Through Adulthood. Cerebral Cortex, 1–14.

Fletcher, S. C. (2018). Computers in Abstraction/Representation Theory. Minds and Machines.

Fodor, J. A. (1968). Psychological Explanation: An Introduction to the Philosophy of Psychology. Random House.

Fodor, J. A. (1975). The Language of Thought. Harvard University Press.

Fodor, J. A. (1981). Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. Representations (pp. 225–256). MIT Press.

Frigg, R., & Hartmann, S. (2018). Models in Science.

Frigg, R., & Nguyen, J. (2018). Scientific Representation.

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

Gould, S. J., & Lewontin, R. C. (1979). The Spandrels of San Marco and the Panglossian Paradigm : A Critique of the Adaptationist Programme. Proceedings of the Royal Society of London, Series B, Biological Sciences, 205(1161), 581–598.

Grothe, B., Pecka, M., & McAlpine, D. (2010). Mechanisms of sound localization in mammals. Physiological Reviews, 90(3), 983–1012.

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

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. Nature, 573(7775), 568–572.

Hardcastle, V. G. (1996). How to Build a Theory in Cognitive Science. SUNY Press.

Harris, C. S. (1965). Perceptual Adaptation to Inverted, Reversed, and Displaced Vision. Psychological Review, 72(6), 419–444.

Hartle, B., & Wilcox, L. M. (2016). Depth magnitude from stereopsis: Assessment techniques and the role of experience. Vision Research, 125, 64–75.

Hillis, D. W. (1998). The Pattern on the Stone: The Simple Ideas that Make Computers Work. Basic Books.

Horiguchi, H., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2012). Human trichromacy revisited.
Proceedings of the First International Conference on Evolutionary Computation and Its Applications, 110(3), E260–E269.

Horowitz, A. (2007). Computation, External Factors, and Cognitive Explanations.
Philosophical Psychology, 20(1), 65–80.

Hubel, D. H., & Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. The Journal of Physiology, 206(2), 419–436.

Huberman, A. D., & Niell, C. M. (2011). What can mice tell us about how vision works?
Trends in Neurosciences, 34(9), 464–473.

Hurvich, L. M., & Jameson, D. (1957). An Opponent-Process Theory of Color Vision.
Psychological Review, 64(6), 384–404.

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

Hutto, D. D., & Myin, E. (2021). Deflating Deflationism about Mental Representation. In J. Smortchkove, K. Dołga, & T. Schlicht (Eds.), What are mental representations? (pp. 79–100). Oxford University Press.

Jacobs, G. H. (2002). Progress Toward Understanding the Evolution of Primate Color Vision.
Evolutionary Anthropology, Suppl 1, 132–135.

Jacobs, G. H. (2008). Primate color vision: A comparative perspective. Visual Neuroscience, 25(5-6), 619–633.

Jacobs, G. H. (2009). Evolution of colour vision in mammals.
Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1531), 2957–2967.

Jacobs, G. H. (2014). The discovery of spectral opponency in visual systems and its impact on understanding the neurobiology of color vision.
Journal of the History of the Neurosciences, 23(3), 287–314.

Jacobs, G. H. (2018). Photopigments and the dimensionality of animal color vision.
Neuroscience and Biobehavioral Reviews, 86, 108–130.

Jacobs, G. H., & Nathans, J. (2007). Response to Comment on "Emergence of Novel Color Vision in Mice Engineered to Express a Human Cone Photopigment". Science, 318(5848), 196.

Jacobs, G. H., & Nathans, J. (2009). The Evolution of Primate Color Vision. Scientific American, April, 56–63.

Jacobs, G. H., Williams, G. A., Cahill, H., & Nathans, J. (2007). Emergence of Novel Color Vision in Mice Engineered to Express a Human Cone. Science, 315(March), 1723–25.

Jameson, K. A., Satalich, T. A., Joe, K. C., Bochko, V. A., Atilano, S. R., & Kenney, M. C. (2020). Human Color Vision and Tetrachromacy. Cambridge University Press.

Jordan, G., Deeb, S. S., Bosten, J. M., & Mollon, J. D. (2010). The dimensionality of color vision in carriers of anomalous trichromacy. Journal of Vision, 10(8), 12–12.

Jordan, G., & Mollon, J. (2019). Tetrachromacy: the mysterious case of extra-ordinary color vision. Current Opinion in Behavioral Sciences, 30, 130–134.

Kanwisher, N., & 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., & Livingstone, M. (n.d.). The neural code for 'face cells' is not face specific. Draft.

Kelber, A., & Jacobs, G. H. (2016). Evolution of Color Vision. In J. Kremers, R. C. Baraas, & N. J. Marshall (Eds.), Human color vision (pp. 317–344). Springer.

Kling, A., Field, G., Brainard, D., & Chichilnisky, E. (2019). Probing Computation in the Primate Visual System at Single-Cone Resolution.pdf. Annual Review of Neuroscience, 42, 169–186.

Kóbor, P., Petykó, Z., Telkes, I., Martin, P. R., & Buzás, P. (2017). Temporal properties of colour opponent receptive fields in the cat lateral geniculate nucleus.
European Journal of Neuroscience, 45(11), 1368–1378.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron, 93(3), 480–490.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience.
Nature Neuroscience, 21(9), 1148–1160.

Kucharski, A. (2016). Post-truth: Study epidemiology of fake news. Nature, 540(7634), 525.

Kwisthout, J., & van Rooij, I. (2020). Computational Resource Demands of a Predictive Bayesian Brain. Computational Brain and Behavior, 3(2), 174–188.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40.

Lazebnik, Y. (2004). Can a biologist fix a radio? Or, what I learned while studying apoptosis. Biochemistry (Moscow), 69(12), 1403–1406.

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. Cortex, 53(1), 60–77.

Lobue, V., & Deloache, J. S. (2011). What's so special about slithering serpents? Children and adults rapidly detect snakes based on their simple features. Visual Cognition, 19(1), 129–143.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". Cognition, 110(2), 248–253.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. Cognition, 99(2), 167–204.

Lombrozo, T., & 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., & Zaitchik, D. (2007). Inferring Design: Evidence of a Preference for Teleological Explanations in Patients With Alzheimer's Disease. 18(11), 999–1006.

Lycan, W. G. (1981). Form, Function, and Feel. The Journal of Philosophy, 78(1), 24–50.

Makous, W. (2007). Comment on "Emergence of novel color vision in mice engineered to express a human cone photopigment". Science, 318(5848).

Mancuso, K., Neitz, M., Hauswirth, W. W., Li, Q., Connor, T. B., Kuchenbecker, J. A., Mauck, M. C., & Neitz, J. (2010). Long-Term Results of Gene Therapy for Red-Green Color Blindness in Monkeys. Invest. Ophthalmol. Vis. Sci., 51(13), 6292.

Mancuso, K., Hauswirth, W. W., Li, Q., Connor, T. B., Kuchenbecker, J. A., Mauck, M. C., & Neitz, J. (2009). Gene therapy for red-green colour blindness in adult primates. Nature, 461, 784–788.

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

Marr, D. (1982). Vision. W.H. Freeman; Company.

Marr, D. (2010). Vision. MIT Press.

Matthews, R. J., & Dresner, E. (2017). Measurement and Computational Skepticism. Nous, 51(4), 832–854.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The Appeal of Parallel Distributed Processing. In D. E. Rumelhart, J. L. McClelland, & P. R. G. The (Eds.), Parallel distributed processing (pp. 3–44). MIT Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133.

Milkowski, M. (2013). Explaining the Computational Mind. MIT Press.

Miller, M. (2014). Minds Online: Teaching Effectively with Technology. Harvard University Press.

Millikan, R. G. (2010). On Knowing the Meaning ; With a Coda on Swampman. Mind, 119(473), 43–81.

Miyauchi, S., Egusa, H., Amagase, M., Sekiyama, K., Imaruoka, T., & Tashiro, T. (2004). Adaptation to left – right reversed vision rapidly activates ipsilateral visual cortex in humans. Journal of Physiology, 98, 207–219.

Mollo, D. C. (2020). Content Pragmatism Defended. Topoi, 39(1), 103–113.

Mollo, D. C. (2021). Deflationary realism: Representation and idealisation in cognitive science. Mind and Language, (July 2020), 1–19.

Mollon, J. D. (1984). Variations of colour vision in a New World primate can be explained by polymorphism of retinal photopigments. Proceedings of the Royal Society of London. Series B, Biological sciences, 222(1228), 373–399.

Morgan, J. L., & Lichtman, J. W. (2013). Why not connectomics? Nature Methods, 10(6), 494–500.

Morrison, J. (2020). Perceptual Variation and Structuralism. Nous, 54(2), 290–326.

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

Nagel, E. (2008). Introduction. In J. A. Boydston (Ed.), The later works of john dewey, volume 12, 1925 - 1953: 1938, logic: The theory of inquiry. Southern Illinois University Press.

Nathans, J. (1999). The Evolution and Physiology of Human Review Color Vision: Insights from Molecular Genetic Studies of Visual Pigments. Neuron, 24, 299–312.

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

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

Neitz, J., & Neitz, M. (2011). The genetics of normal and defective color vision. Vision Research, 51(7), 633–651.

Neitz, J., & Neitz, M. (2017). Evolution of the circuitry for conscious color vision in primates. Eye (Basingstoke), 31(2), 286–300.

Neitz, M., & Neitz, J. (2014). Curing Color Blindness—Mice and Nonhuman Primates. Cold Spring Harbor Perspectives in Medicine, 4, 1–13.

Nersessian, N. J. (2002). The cognitive basis of model-based reasoning in science. In P. Carruthers, S. Stich, & M. Siegal (Eds.), The cognitive basis of science (pp. 133–153). Cambridge University Press.

Neumann, J. v. (1958). The Computer and the Brain. Yale University Press.

Nityananda, V., & Read, J. C. (2017). Stereopsis in animals: Evolution, function and mechanisms. Journal of Experimental Biology, 220(14), 2502–2512.

Niv, Y. (2020). On the Primacy of Behavioral Research for Understanding the Brain. In A. J. Lerner, S. Cullen, & S.-J. Leslie (Eds.), Current controversies in philosophy of cognitive science (pp. 134–149). Routledge.

Paul, L. (2014). Transformative Experience. Oxford University Press.

Peacocke, C. (1994). Content, Computation and Externalism. Mind & Language, 9(3), 303–335.

Peacocke, C. (1999). Computation as involving content: A response to Egan. Mind and Language, 14(2), 195–202.

Piccinini, G. (2008). Computation Without Representation. Philosophical Studies, 137(2), 205–241.

Piccinini, G. (2015). Physical Computation: A Mechanistic Account. Oxford University Press.

Piccinini, G., & Shagrir, O. (2014). Foundations of computational neuroscience. Current Opinion in Neurobiology, 25, 25–30.

Platt, M. L., & Ghazanfar, A. A. (Eds.). (2010). Primate Neuroethology. Oxford University Press.

Polanyi, M. (1966). The Tacit Dimension. Doubleday.

Poldrack, R. A. (2010). Mapping mental function to brain structure: How can cognitive neuroimaging succeed? Perspectives in Psychological Science, 5(5), 753–761.

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

Putnam, H. (1991). Representation and Reality. MIT Press.

Pylyshyn, Z. W. (1984). Computation and Cognition. MIT Press.

Pylyshyn, Z. W. (1993). Computing in Cognitive Science. In M. I. Posner (Ed.), Foundations of cognitive science (pp. 49–92). MIT Press.

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

Ramsey, W. M. (2021). Defending Representation Realism. In J. Smortchkove, K. Dołga, & T. Schlicht (Eds.), What are mental representations? (pp. 55–78). Oxford University Press.

Rescorla, M. (2013). Against Structuralist Theories of Computational Implementation. The British Journal for the Philosophy of Science, 64, 681–707.

Rescorla, M. (2017). The Computational Theory of Mind.

Rhodes, G., Byatt, G., Michie, P. T., & Puce, A. (2004). Is the Fusiform Face Area Specialized for Faces, Individuation, or Expert Individuation? Journal of Cognitive Neuroscience, 16(2), 189–203.

Richards, B. A., & Lillicrap, T. P. (2022). The Brain-Computer Metaphor Debate Is Useless: A Matter of Semantics. Frontiers in Computer Science, 4(February), 1–8.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. Nature Neuroscience, 22(11), 1761–1770.

Richmond, A. (n.d.-a). Computational Externalism. Forthcoming.

Richmond, A. (n.d.-b). How Computation Explains. Forthcoming.

Richmond, A. (n.d.-c). What is a Theory of Neural Representation For? Forthcoming.

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

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. 8(12), 1647–1651.

Sabesan, R., Schmidt, B. P., Tuten, W. S., & Roorda, A. (2016). The elementary representation of spatial and color vision in the human retina. Science Advances, 2(9).

Samuels, R. (2019). Classical Computational Models. In M. Sprevak & M. Colombo (Eds.), The routledge handbook of the computational mind (pp. 103–119). Routledge.

Sánchez, V. G. (n.d.). What Bayesian Angels have to do with Human Cognition. Forthcoming.

Scheutz, M. (2012). What it is not to Implement a Computation: A Critical Analysis of Chalmers' Notion of Implementation. Journal of Cognitive Science, 13(1), 75–106.

Schneider, S. (2019). Artificial You: AI and the Future of Your Mind. Princeton University Press.

Searle, J. R. (1980). Minds , brains , and programs. The Behavioral and Brain Sciences, 3, 417–457.

Searle, J. R. (1992). The Rediscovery of Mind. MIT Press.

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational Neuroscience. Science, 241(4871), 1299–1306.

Seung, S. (2012). Connectome. Mariner.

Shadmehr, R., & Wise, S. (2005). The Computational Neurobiology of Reaching and Pointing. MIT Press.

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

Shagrir, O. (2018). In defense of the semantic view of computation. Synthese, (January).

Shagrir, O. (2022). The Nature of Physical Computation. Oxford University Press.

Shapley, R. (2009). Gene Therapy in Color. Nature, 461, 737–738.

Shea, N. (2013). Naturalising Representational Content. Philosophy Com, 8(5), 496–509.

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

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

Shepherd, G. M. (1991). Foundations of the Neuron Doctrine. Oxford University Press.

Shepherd, S. V., & Platt, M. L. (2010). Neuroethology of Attention in Primates. In M. L. Platt & A. A. Ghazanfar (Eds.), Primate neuroethology (pp. 525–549). Oxford University Press.

Shevell, S. K., & Martin, P. R. (2017). Color opponency: tutorial. Journal of the Optical Society of America, A, 34(7), 1099–1108.

Simon, H. A., & Newell, A. (1973). Human Problem Solving: The State of the Theory in 1970. American Psychologist, 26(2), 145–159.

Smith, B. C. (1999). Computation. In R. A. Wilson & F. C. Keil (Eds.), The mit encyclopedia of the cognitive sciences (pp. 153–155). MIT Press.

Sower, V. E., Duffy, J. A., & Kohers, G. (2008). Ferrari's Formula One Handovers and Handovers From Surgery to Intensive Care. The American Society for Quality, (August), 1–5.

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

Sprevak, M. (2019). Triviality arguments about computational implementation. The routledge handbook of the computational mind (pp. 175–191). Routledge.

Sterelny, K. (1990). The Representational Theory of Mind: An Introduction. Basil Blackwell Inc.

Stich, S. (2010). Autonomous Psychology and the Belief-Desire Thesis. Collected papers: Mind and language (pp. 53–70). Oxford University Press.

Strawson, P. F. (1974). Freedom and Resentment. Freedom and resentment and other essays (pp. 1–28). Routledge.

Sun, D., Lv, J., & Waller, S. T. (2011). In-depth analysis of traffic congestion using computational fluid dynamics (CFD) modeling method. Journal of Modern Transportation, 19(1), 58–67.

Sun, R. (2008). Introduction to Computational Cognitive Modeling. In R. Sun (Ed.), The cambridge handbook of computational psychology (pp. 3–20). Cambridge University Press.

Thomson, E., & Piccinini, G. (2018). Neural Representations Observed. Minds and Machines, 28(1), 191–235.

Thoreson, W. B., & Dacey, D. M. (2019). Diverse Cell Types, Circuits, and Mechanisms For Color Vision in The Vertebrate Retina. Physiol Rev, 99, 1527–1573.

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.

Wachtler, T., Dohrmann, U., & Hertel, R. (2004). Modeling color percepts of dichromats. Vision Research, 44, 2843–2855.

Wachtler, T., & Wehrhahn, C. (2016). Computational Modeling of Color Vision. In J. Kremers, R. C. Baraas, & N. J. Marshall (Eds.), Human color vision (pp. 243–268). Springer.

Waismann, F. (1968). Verifiability. In R. Harré (Ed.), How i see philosophy. Palgrave Macmillan.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience, 19(3), 356–365.

Zele, A. J., & Cao, D. (2015). Vision under mesopic and scotopic illumination. Frontiers in Psychology, 5, 1–15.