COMPUTATIONAL EXTERNALISM

Andrew Richmond

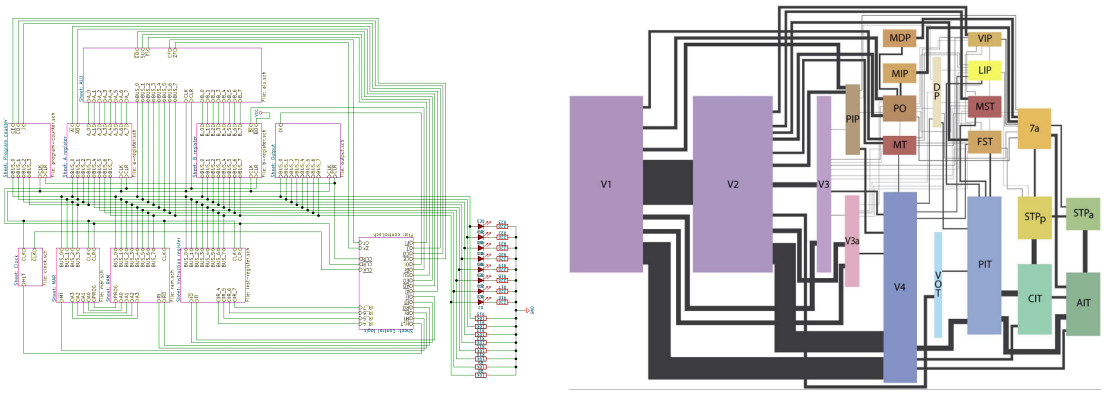
ABSTRACT

I argue that the brain does not have its computational structure intrinsically, but only in relation to its environment. I support this view (*externalism*) with a case study in the neuroscience and evolutionary biology of color vision, showing that which aspects of the brain's causal structure rise to the level of computation — which features of its causal structure count as part of its functional structure or "wiring diagram" — depends on its environment. I show that this version of externalism helps answer some pressing methodological questions in neuroscience and explainable AI. Along the way I connect some traditional debates about externalism to contemporary cognitive science, and demonstrate the promise of a deflationary approach to cognitive scientific explanation.

1 Introduction

Neuroscience has begun to develop extremely detailed causal models of the brain. That project has been so successful that we are beginning to see connectome maps of whole cortical areas, and it has been so influential that a map of every neuron and synapse in the human brain has seemed, at least to a few mavericks and funding agencies, just off the horizon (Markram 2006; Naddaf 2023). But this progress comes alongside renewed debate about how much neuroscience can learn from brain data alone, without detailed and well-theorized data about behavior and its relation to the environment (Krakauer et al. 2017; Niv 2020). To understand the computational or functional structure of the brain, is it enough to understand its internal causal structure, or do we need to carefully study the structure and dynamics of its surroundings too?

For neuroscience, these are methodological questions. But they are tied to an older philosophical one: are the brain's functional or computational properties *intrinsic* to it, or do they depend on things outside the brain (e.g., Peacocke 1994; Egan 1995)? By comparison, the schematic of a typical computer (see Figure 1a) seems to describe an intrinsic property of the computer. It looks



(a) (b)

Figure 1: Schematics for (a) an 8-bit computer, from Ben Eater (https://eater.net/8bit/schematics), and (b) the primate visual system (Wallisch and Movshon 2008).

for all the world like someone just laid out the computer's innards and traced them. It might stretch the metaphor to think of computational models as *tracing* the brain, but it is natural to think that they too capture something intrinsic to it: a structure that exists in the brain independently of its environment, the way a memory bus exists independently of the computer's environment (compare Figures 1a & b). Internalists think this comparison holds up. Externalists don't. Externalists think the brain's computational properties reflect a relationship between brain and environment, rather than a feature of the brain considered in isolation.

In §2 I will describe a series of methodological issues in neuroscience, all concerning the relevance of the environment to our computational understanding of the brain. I will suggest that the traditional philosophical debate over externalism can help answer those questions, though we have to update that debate in a few important ways. In §3 I will argue for externalism using a case study from the neuroscience and evolutionary biology of color vision. And in §4 I will apply externalism to the aforementioned methodological issues. I will conclude by reflecting on the *philosophical* methodology implicit in my argument, and how it differs from attempts to learn about computational explanation via the nature or definition of computation.

2 What is externalism, and why does it matter?

Externalists hold that the brain's computational properties are not intrinsic to the brain, but reflect a relationship between brain and environment. My goal, in this section, will be to refine this thesis so that it connects straightforwardly to some methodological debates that are ongoing in neuroscience, and likely on the horizon for explainable AI. I'll begin by outlining those debates.

2.1 Methodological debates concerning brain and environment

First, and most straightforwardly, Krakauer et al. (2017) revived some perennial questions about how we investigate the computational structure of the brain. Specifically, they asked whether, to reveal that structure, we can just interrogate the brain itself, or if we must also study its environment and its behavior in that environment. As the authors put it in an analogy with computer science: neuroscience seeks to understand the processes governing behavior, and the "core question ... is whether the processes governing behavior are best inferred from examination of the processors" themselves (Krakauer et al. 2017, p. 480; cf. Jonas and Kording 2017).

Second, *connectomics* has caused a great deal of both excitement and skepticism in neuroscience. Connectomics aims, eventually, to map every neuron and synapse in the brain. There are already good connectome maps for simple organisms like C. elegans and drosophila, some progress on the mouse connectome, and at least optimism about a connectome for areas of the human brain (see Elam et al. 2021, and their references). And while there is a growing appreciation in neuroscience for the power of this approach, it remains unclear what it can tell us about the *computational* structure of the brain, as opposed to its fine-grained *causal* structure. The prospect of a complete map of the brain's neurons and synapses raises Krakauer et al.'s question in a particularly striking way: if we had all the information we could possibly want about the brain itself, what would remain for us to learn from the environment outside the brain? Would considering that environment help us change or improve our connectome maps? Would it merely add context showing how the connectome relates to the environment? Or would it push us develop *alternative* maps to capture the brain's computational — not just causal — structure? Given the massive funding that connectomics projects have received, this is quite literally a billion-dollar question.

Third, consider the ongoing debate over naturalistic and ecologically valid stimuli. It is widely agreed that, on *some* definition, naturalistic stimuli are preferable to non-naturalistic ones. But reasons vary. Some argue that the proper object of study for neuroscience is the organism-environment system (Chemero 2009), in which case it would be worse than useless to study the organism in the wrong environment. Some hold that naturalistic stimuli evoke "natural behaviors" that are robust and species-typical and, therefore, the proper target for neuroscience (Krakauer et al. 2017, p. 481; but see Bolt et al. 2018). Others suggest that only naturalistic stimuli will reveal the way that our brains are "tuned to" the dynamics of our natural environments (Sonkusare et al. 2019, p. 300). To determine whether and how experimental environments should approximate real environments, we need to know whether and in what way the environment matters for neuroscience.

In each of these debates, the question is *whether* and *how* the environment matters for our understanding of the brain's computational structure. And each debate will likely arise in machine

learning as well as neuroscience. The fields of *explainable AI* (XAI) and *machine behavior* seek to understand complex computational systems, the latter especially using strategies and concepts from the life sciences (Rahwan et al. 2019; Merel et al. 2019). Machine learning has its own notions of computation, capturing both model architecture (e.g., HCNNs differ computationally from transformers) and the fine-grained structure of nodes and their connections (e.g., convolutional filters will differ computationally depending on their weights). But neuroscience has an intermediate notion of representation, capturing (or idealizing) the way information flows through a complex system (Marr 1982; Kriegeskorte and Diedrichsen 2019; Doerig et al. 2023), and it is likely that XAI and machine behavior will benefit from this intermediate level of description. To whatever extent these fields do take up a similar notion of computation, they will also have methodological questions that call for clarity about how computational structure is to be sought: by investigating a network's own properties, or those of its environment as well. I'll return to methodological questions in §4. But next, I will formulate and clarify the version of externalism that will help us intervene on them.

2.2 Refining externalism: structure and dependence

These methodological debates do not, like philosophical externalism, ask what *makes it the case* that the brain has the computational properties it has. They ask what sources of *evidence* we should attend to if we want to discern the brain's computational properties. But the former question can help us answer the latter. If externalism is wrong, and the brain has its computational properties intrinsically, then we may be able to discern those properties by studying the brain itself (contra Krakauer et al. 2017). We should expect a sufficiently advanced connectomics to reveal the brain's computational properties, as some proponents of connectomics suggest (Schneider 2019, p. 115-116, Seung 2012). And naturalistic stimuli should matter to neuroscience only insofar as, by driving the brain in particular ways, they reveal more about its intrinsic features.

But if externalism is right and the brain has its computational structure only extrinsically, then none of those strategies are plausible. We should not expect to glean the brain's computational properties by studying the brain alone — we also have to study the features of the environment on which those properties depend. Even the extreme detail connectomics promises will not reveal computational properties; as the externalists of old might have put it, computation just ain't in the connectome (Putnam 1975). And our choice of stimuli must depend on more than their ability to reveal the brain's intrinsic properties: stimuli with the features that (partly) determine the brain's computational properties will be more informative than stimuli *without* those features.

There are a few things about externalism we have to make clear, in order to draw any of these conclusions. As I described externalism, it claims that the brain's computational properties are not intrinsic to the brain: they depend partly on its environment. I want to clarify what I mean by

computational properties, and then say more explicitly what it would mean for them to *depend on* the environment.

For computational externalism to even be relevant to the methodological debates I've discussed, the computational properties at issue will have to be the same ones that are at issue in the methodological debates. And those debates concern the brain's *structure* (cf. Cummins 1991; Egan 2010, 2014; Pylyshyn 1993). Computational descriptions of the brain are process models, or causal models, of the brain. They aren't *just any* causal models. The brain is, famously, a "mess" of causal detail (Crick 1979), and we talk about a system's computational structure to idealize, coarse-grain, and abstract away from its fine-grained causal structure in a distinctive way, using the resources that the notion of computation introduces (Richmond n.d.*a*). It is this high-level causal structure that computational models are supposed to capture, and that the methodological debates I discussed are about. Externalism, as I will defend it, says that *this structure* is not intrinsic to the brain.

I mention this because the philosophical tradition has tended to focus on other computational properties, especially the correct *categorization* of computational structures (Peacocke 1992, 1999; Egan 1999). The most common externalist strategy has been to argue that computational descriptions must refer to the representations a computation is defined over, like representations of edges in visual computations (Fletcher 2018; Peacocke 1994, 1999; Piccinini 2008, 2015; Piccinini and Shagrir 2014; Rescorla 2013; Shagrir 2001, 2020, 2022; Shea 2013; Sprevak 2010). Then, if the identity of a representation depends on its environment (e.g., a representation of edges might, in an alien environment, be a representations of shadows), our description of computations defined over that representation will depend on the environment too. The traditional debate is about whether a given computational structure, defined over different representations, counts as *the same* computation. Externalists stress that computations over different representations have different counter-factual implications (Peacocke 1992) and explain different facts (Peacocke 1994, 1999). That suggests we should distinguish between the computations. Internalists argue that it is more fruitful, or more faithful to cognitive scientific practice, to classify computations based solely on their causal or syntactic structure, regardless of the representations they operate on (Egan 2014, 1992).

These debates are important, but they concern how computational structures are *classified*, not how they are *determined*. And the methodological debates I want to intervene on don't ask how a given computational structure should be classified, but how that computational structure should be discovered in the first place. The kind of externalism that will be relevant will therefore say that what computational structure the brain has — not just how we classify that structure — depends on the environment.

There are, to my knowledge, only four philosophers who endorse this kind of externalism, which I'll call *structural externalism*: Bontly (1998), Horowitz (2007), Shagrir (2001, 2020), and

Shea (2013).[1] Of these four, all but Bontly argue from a representational account of computation, and Bontly invokes a teleological account instead. That is, these authors follow the argumentative strategy of traditional externalism. They just extend the argument to say that changing the representations (or teleological functions) a computation is defined over doesn't just force us to classify the computation differently, it changes our understanding of the computation's structure as well, e.g., by grouping different sets of physical states together as the 'same' computational state (Shagrir 2001).

I won't take that path, for a few reasons. First, the arguments in question tend to rely on toy examples rather than cases of real scientific reasoning. Toy examples can be informative, but a main contribution of this paper is to bring the argument for externalism beyond them. Second, teleological and (especially) representational accounts of computation have become increasingly contentious (Piccinini 2008), and the case study I consider will show that we don't need them anyways. And finally, by divorcing externalism from representation we can refute some common assumptions: that a narrow account of content (as in Butler 1998; Shagrir 2001) or a non-representational account of computation (as in Egan 1991, 1995, 1999, 2010, 2014) would establish internalism. If externalism can be demonstrated without representationalism, both assumptions are incorrect

Moving on to the second point of clarification, what would it be for a system's computational structure to *depend on* its environment? Egan describes two popular ways of cashing out the idea (though, again, with a focus on classification, or "taxonomy," rather than structure):

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

The *reference* and *supervenience* definitions are generally taken to be equivalent. In §3.2 I'll argue that they are not. In the meantime, I'll work with the reference definition because the supervenience definition won't apply to my case study, which involves changes to the internal structure of the brain.

2.3 A toy example of structural externalism

Apart from the structural externalists I mentioned above, there is a broad consensus against the view. After all, I've described computational structure as a kind of *causal* structure, and a system's causal

[1] They have tended to call the view *syntactic* or *vehicle* externalism (e.g., Shea 2013; Shagrir 2001), but both names reflect representationalist concerns — externalism about syntax as opposed to semantics, and vehicles as opposed to their contents — and I don't want to introduce that baggage here.

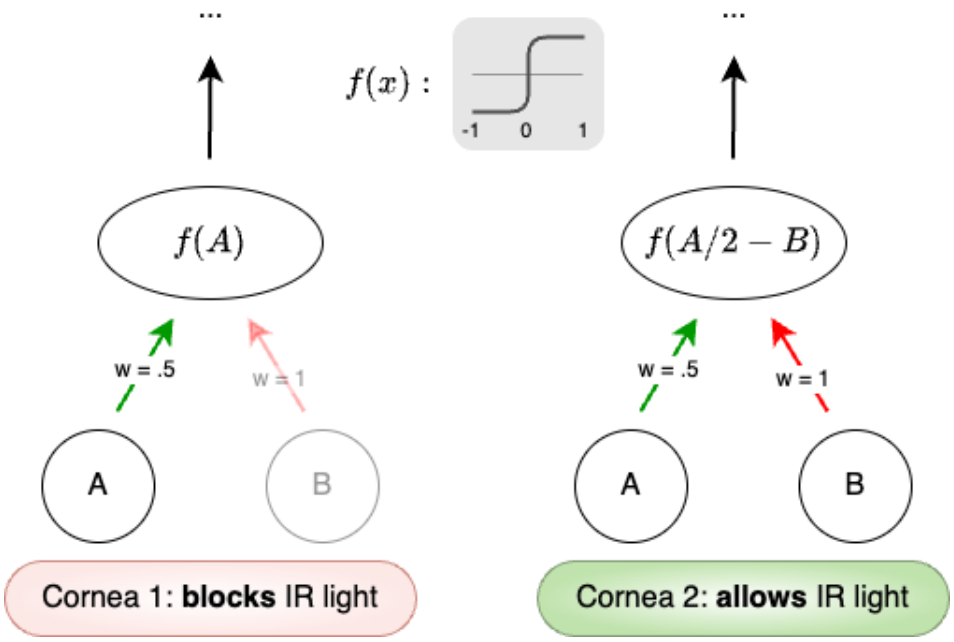


Figure 2: A network with a simple computational structure, before (left) and after (right) a mutation to its cornea.

structure seems to be a paradigmatically intrinsic property (Shea 2013). So I want to illustrate structural externalism with a toy example before moving on to the argument.

The point of this section is not to argue for externalism about the toy example. If we want to understand the notion of computation *in cognitive science*, it would be of limited value to prove externalism about a system that is not studied by cognitive science, with our investigation unconstrained by the typical explanatory context of cognitive science (pace Shagrir 2001; Shea 2013). Rather, I'll use the toy example to show that structural externalism is at least within the realm of possibility, and to clarify the basic structure of the case study to come.

So, consider a simple visual network like the one on the left side of Figure 2. It has two input nodes, A and B, which are excited by light. A is excited by light around 400nm; B is excited by infrared light around 900nm. Each node passes on its activity, with connection strengths denoted by *w*'s, to an output node. 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 is one complication to this toy system: it has a cornea that (like ours) filters light before it reaches the input nodes. Specifically, in this organism, *the cornea filters out all infrared light*. The filtering means that B will never be active. Because of that, the output node appears to have a very simple job: it computes the threshold function for A, $f(A)$. B is irrelevant to the functioning of the

system. Maybe it is a spandrel; maybe it is useful for some purpose unrelated to this computation. The point is that B does not figure into the system's *computational* structure, though it is a part of the system's causal structure. To reiterate: I'm not giving an argument for this. We don't know much about this system, or the goals and constraints that would govern an investigation of its computational structure. In §3, when we come to the actual case study, we'll have both in spades. For now, the toy example is just illustrating the shape than an argument for externalism would take. The first step of that argument would be to establish that a system has a certain computational structure, which includes and excludes parts of its finer-grained causal structure.

But now imagine this little system evolves a new cornea, which allows infrared light through. The new network is illustrated on the right side of Figure 2. Now B *will* drive the output node. Note that we haven't changed the circuits whose computational structure we're describing. We've just intervened on the cornea. But now it would clearly be wrong to say that the output node is just computing $f(A)$. B has an independent effect on the output node, which now appears to be computing the threshold function on a weighted combination of A and B: $f(A/2 - B)$. Suddenly we have a more complicated computational structure. In short, it is plausible that the computational structure of the second system would include features that were not part of the computational structure of the first system. Then the computational structure of the system would have changed without any change to the computing circuits — just the cornea.

Assuming I've gotten the computational structure of these two systems right, what follows? As I put it above, an internalist is committed to the computational structure of a system depending entirely on features internal to that system. The internalist is therefore committed to there being a reason or ground for the difference in computational structure, where that reason or ground concerns only features of the systems themselves. And one might be skeptical that the internalist can provide this reason. It's not enough to say there was a change to the cornea, of course: not just any change to the cornea changes the system's computational structure. So what *about* the change to the cornea grounds the computational difference? If we can answer that question without referring to the environment, internalism is a live option. If not, and we have to refer to the environment, we're stuck with externalism.[2]

It would be too high a bar to demand an externalist to show that an internalist *could not* devise an appropriate reason for the computational re-description. It should be enough for the externalist to show that the resources available to the internalist are scant and unpromising, while the resources available to the externalist are abundant and well-suited to the task of capturing differences in

[2]We can already see how the supervenience and reference definitions come apart. If the internalist can't give plausible grounds for the computational difference that results from a change to the cornea, internalism fails according to the reference definition. But, because there *was* a change to the internal structure of the system — specifically, to the system's cornea — internalism will *not* have failed according to the supervenience definition.

computational structure (and in a way that allows us to intervene on the methodological debates that are ultimately at issue). That would be a strong argument for externalism, and it is the argument I'll make in the next section.

3 Externalism in the study of color vision

Following the logic of the toy example, §3.1 will show that two organisms, at different stages of evolutionary development, have different computational structures. And §3.2 will argue that there is likely no reason or ground for this difference that appeals only to the organisms' internal features.

3.1 Case study: color vision

The main finding I'll rely on is this: primates' evolution from dichromacy to trichromacy likely occurred without any relevant changes to circuitry beyond the retinal mosaic. But dichromacy and trichromacy require different *computational* structures in post-mosaic circuits. This is analogous to the toy example: this time it is the retinal mosaic, rather than the cornea, that changes. But again we have a peripheral change to the system's causal structure, which calls for significant changes to the computational structure of other, less peripheral parts of the system. I'll start by discussing the computations involved in trichromatic color vision. Then I'll summarize the work in evolutionary biology that supports the finding I just described.

Humans, along with the rest of the Old World primates, are behavioral trichromats, meaning that we extract three dimensions of color experience from visual stimuli (Jacobs 2002, 2009; Jacobs and Nathans 2009). That is, there are three monochromatic lights, i.e. three wavelengths, such that for any color we see, that color is indistinguishable from some weighted combination of the 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*.[3]

How do we extract these three dimensions of color experience? Kelber and Jacobs summarize the requirements of color vision 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 2016, 106)

[3]The possible tetrachromacy of humans in mesopic light conditions (Zele and 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).

Primates have a mosaic of cone cells in the first layer of the retina, expressing different *opsins*. Each opsin has a unique *wavelength sensitivity profile*, describing its response to different wavelengths of light. In humans and other Old World primates, cone cells express either an S, M, or L opsin, named for their peak sensitivities to short-, medium-, and long-wavelength light. Step (1) consists of those cone cells' different responses to an incoming wavelength profile.

In step (2), some mechanism has to collect these responses and determine a single color percept. In primates, that is accomplished by a process that begins with circuits performing some 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 and 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 ongoing debate about the 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 (Neitz and Neitz 2011). But these debates are inessential for my purposes. What matters is that color vision is performed, in humans and our close relatives, partly by mechanisms implementing computations defined over the activation levels of the three types of photoreceptors — everything I'll say can be translated to different models of those computations.

Much of this is widely known — philosophers have been interested in the computational structure of color vision for some time (e.g., Hardin 1988). What isn't so familiar in philosophical circles is the evolutionary biology of color vision. 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, meaning that males of the species get just one copy of each, while females get two. Our ancestors, however, had *just one* of the longer-wavelength opsins: the M opsin.[4] Males received one copy of the M opsin gene on their X chromosome, while females received two copies, one on each X chromosome. Half of a female's cone cells would have expressed one chromosome and half the other,[5] but since both chromosomes had the same M opsin gene, the result was a two-opsin retinal mosaic just like a male's, made up only of S and M cones. These primates were dichromats per the definition above, only able to extract two dimensions of color from stimuli (Jacobs 2002, 2009; Jacobs and Nathans 2009; Nathans 1999).

[4]It's not actually clear whether the M or L opsin came first, or in what order the alleles of the M and L opsin genes evolved. I'll call the original the M opsin for the sake of simplicity, but nothing hinges on this decision.

[5]A random process determines which X chromosome is expressed in each cell (Jacobs 2008; Neitz and Neitz 2011).

The *three-opsin mosaic* is the star of this story, and it came about in two separate ways. In Old World primates, it appears that a *gene duplication* resulted in an X chromosome with the opsin gene at two separate locations. A later mutation at one of those locations resulted in a gene encoding a new opsin: the L opsin.[6] The first primate with that mutation would have had a three-opsin retinal mosaic.[7] In New World monkeys a similar change occurred, but without the duplication event: each animal's X chromosome has just one opsin gene, but mutations have created multiple alleles of that gene in the population.[8] In these species males have a two-opsin mosaic, with just the S opsin and a single longer-wavelength-sensitive opsin on their single X chromosome.[9] But females lucky enough to receive different alleles of the opsin gene on their two X chromosomes have a three-opsin retinal mosaic, with half of their cells expressing one chromosome and half the other (Dulai et al. 1999; Jacobs 2009; Jacobs and Nathans 2009).

The finding I mentioned above is that the evolution from dichromacy to trichromacy seems to have occurred *immediately* on the introduction of the third type of photoreceptor cell. That is, the first primate to express S, M, *and* L opsins in its retinal mosaic would have had a new dimension of color vision,[10] with no changes to post-retinal circuits (Conway et al. 2010; Huberman and Niell 2011; Jacobs 2008, 2009; Jacobs and Nathans 2009; Kóbor et al. 2017; Mancuso et al. 2009, 2010; Mollon 1984; Neitz and Neitz 2011, 2014; Shapley 2009; Wachtler et al. 2004; Wachtler and Wehrhahn 2016; Chang et al. 2013). Or at least: no changes specifically related to trichromacy. It is an idealization to assume that *any* two organisms have the same post-mosaic circuits, since those circuits are generated partly randomly. But differences that could not plausibly ground a difference in computational structure are irrelevant, and the random variation that exists between all primates (cross-cutting dichromats, trichromats, and monochromats) could not ground a distinction between dichromats and trichromats. So when I, or the scientists I quote, say "no changes to the circuits" we mean no *relevant* changes; we are setting aside all purely random differences in neural circuitry.

[6] It's possible that the mutation came first, in which case the three-opsin mosaic in Old World primates would have developed like it did in the New World monkeys I discuss below, but with an additional, *later* gene duplication putting two opsin genes on the same X chromosome.

[7] A random process also determines which of a given X chromosome's opsin genes is expressed in a cell (Jacobs 2008; Neitz and Neitz 2011).

[8] The howler monkey is the one known exception. It appears to have independently undergone a similar process to the Old World primates (Jacobs 2002, 2009).

[9] Which allele a male has doesn't seem to make a difference to the computations involved in color vision, though it does affect some of their discriminatory abilities (Neitz and Neitz 2011, p. 639).

[10] 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, including Gerald Jacobs (personal communication), who performed one of the experiments I discuss below. He notes that in his experiments on mice, the new dimension of color vision is confined to a portion of the wavelength spectrum where mice *may* have been effectively monochromatic. But the new cone type nonetheless increased the dimensionality of color vision along that portion of the wavelength spectrum without altering post-mosaic 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.

Arguments for the immediate development of trichromacy appeal to two main lines of evidence. First, it is widely agreed that there must have been some selective advantage in order for the new opsin gene alleles, and especially the X chromosome with two opsin gene locations, to spread through the population (Jacobs and Nathans 2009; Mancuso et al. 2010). That advantage may have been the new abilities that trichromacy affords: to discriminate ripe fruit from foliage, or to identify skin tone and socially important features of conspecifics (Dulai et al. 1999; Jacobs 2009; Jacobs and Nathans 2009). For familiar reasons (Gould and Lewontin 1979), I won't lean heavily on this adaptationist argument, but it is taken seriously in this literature.

The second line of evidence is more compelling: the evolutionary step in question 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. Jacobs and colleagues inserted a new long-wavelength opsin allele into "knock-in" mice's X chromosomes, and at adulthood heterozygote females — those 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).[11]

Another experiment was performed with adult male squirrel monkeys (Mancuso et al. 2009). Squirrel monkeys, like the rest of the New World monkeys, have the sex-specific trichromacy I described above: there are different alleles of the X chromosome in their population, but there is still only room for one opsin gene on the X chromosome. Male squirrel monkeys therefore have a two-opsin mosaic, and are dichromats. Females who receive different alleles on their two X chromosomes have a three-opsin mosaic, and are trichromats. Mancuso et al. (2009) injected a virus into male squirrel monkeys' retinas carrying an allele of the M opsin gene, 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, a new type of cone cell transforms a dichromat into a trichromat.

There is no general rule that organisms with n cone types have n-dimensional color vision (Jordan et al. 2010; Neitz and Neitz 2014). The organism needs a way to take advantage of the diversity of its photoreceptors — that was step (2) in the account of color vision I began with. According to our best understanding of color vision, organisms take advantage of their different cone types with post-mosaic structures that perform computations on them. And the post-mosaic computations supporting trichromacy and dichromacy are different. This should be intuitive: the trichromat's post-mosaic circuits perform, e.g., L – M computations, and the dichromats in question

[11] A small group of dissenters argue the mice did not develop a new dimension of color vision, just new sensitivities to texture and illumination that can be mistaken for trichromacy (Makous 2007; Cornelissen and Brenner 2015). This is less plausible in the other experiment I'll discuss, so I'll set it aside.

have no L cones for that computation to be defined over. This point is borne out by current work on the evolution of trichromacy, which I'll return to momentarily. But first I want to address a more pressing concern.

My argument will depend on the claim that a change in the retinal mosaic *alone* was sufficient to turn dichromats into trichromats. But might there also have been relevant changes to the post-retinal circuits, allowing them to take advantage of the new cone types (Nathans 1999; Wachtler et al. 2004)? After all, the mice had an entire adolescent period with the new photoreceptor mosaic — their post-mosaic mechanisms may have developed to take advantage of that mosaic (Neitz and Neitz 2011; Jacobs and Nathans 2007). This is a tempting idea, but it is less plausible than it appears. In the first place, there is no obvious reason that developmental plasticity would be able to accommodate three types of cone cell but not four or five (cf. Neitz and Neitz 2011, p. 635). And consider the many human females with four or five cone types who are not tetra- or pentachromatic. The evidence that even *some* human females are even *tetra*chromatic is controversial and inconclusive (Jordan et al. 2010; Jordan and Mollon 2019). If plasticity is responsible for the new dimension of color vision, we need an account of its apparently arbitrary dimensional limits.

Moreover, the experiment on squirrel monkeys was performed long after critical development periods — when the brain is most plastic — had ended (Feldmann et al. 2018; Hubel and Wiesel 1970), and the monkeys developed trichromacy "just as levels of transgene expression [the presence of the new opsin] became robust" (Neitz and Neitz 2014). It is widely agreed that this is too soon for the new visual capacity to be due to changes in early retinal circuits, where the computations at issue are performed (Shapley 2009; Conway et al. 2010; Mancuso et al. 2009, 2010). Though the evidence is not conclusive, the consensus, regarding the experiments on squirrel monkeys at least, is that "no rewiring or new circuitry was associated with the acquisition of red–green color vision" (Neitz and Neitz 2011, p. 642). Or, as it is expressed by Chang et al. (2013):

> This rapid emergence of a new dimension in color space in former dichromats may be due to the formation of novel circuits; however, it is more likely that preexisting, cone-type unselective retinal microcircuits are able to extract (new) chromatic information. (Chang et al. 2013, p. 559)

So, with the squirrel monkeys at least, we have a situation like the toy example. A change in the photoreceptor mosaic, with no corresponding change to post-mosaic circuits, nonetheless changes the computational structure of those post-mosaic circuits — at least insofar as trichromacy calls for a different computational structure than dichromacy. A moment ago I postponed a discussion of that point: that the trichromat's color vision really does call for a different computational structure than the dichromat's. I want to conclude this subsection by returning to that point, since we might

wonder: do the squirrel monkeys really have a different computational structure before and after the injection of the new opsin gene?

A negative answer would mean accepting that the monkeys were performing L – M computations all along, despite the fact that they had no L cones, and the most those computations could do would be to subtract M cone activity from M cone activity. I won't argue that this is impossible, or a completely useless computation.[12] But I'm not aware of any work on dichromacy that takes this to be a part of any dichromat's computational structure, and it is clearly inappropriate to *impose* a commitment to the M – M computation on cognitive science — especially when our goal is to understand computation as that notion is used in cognitive science. Let me briefly give two examples of work on the development of trichromacy that endorse the idea that, although the causal structure of the post-mosaic circuits did not change, their computational structure did.

First, Chang et al. (2013) note that L vs M opponent retinal ganglion cells (the cells responsible for L – M or M – L computations) gain their opponency through their center-surround structure: they receive excitatory input from a single cone cell in the center of their receptive field, and inhibitory input from the many cones surrounding it. The authors suggest that these cells originally supported shape perception by detecting luminance contrasts through a center vs surround computation (p. 566; see also Shapley 2009). But when a new cone type is introduced, the surround of these ganglion cells' receptive fields will now contain both M and L cells, with the center containing *either* an M or L cell. That means they will automatically be extracting spectral information: a cell receiving excitatory input from one L cell (in the center) and inhibitory input from a collection of M and L cells (in the surround) will be performing a weighted L – M computation. It is important that in this computation *neither L nor M refers to the center cone alone*, but to a weighted average of the center cone and the surround cones of the same type (Chang et al. 2013). The new computation is not a center vs surround computation.[13]

Neitz and Neitz (2011) and Shapley (2009) instead suggest that the retinal ganglion cells responsible for L vs M computations originally performed color vision computations — just a different set of computations than they perform in the trichromat. Specifically, they suggest that these cells were originally part of an S vs M opponent system, and the introduction of L cones 'split' that system in two so that instead of computing M – S (with an M cell in the center), they now compute either M – (S + L) (with an M cone in the center) or L – (S + M) (with an L cone in the

[12] John Morrison (in conversation) points out that it could be used to measure the variance in M cone activity.

[13] The center-surround *causal* structure remains, but is no longer an essential feature of the computation. It just gives the center cone a greater weight than any of the surround cones. It would be accurate to describe the circuit as doing something like $L_{center} - (L_{surround} + M_{surround})$, but in the actual computational descriptions of this circuit these center-surround details are set aside, with the surround L cones serving just to lower the weight of the L term in a weighted $L_{entire\ field} - M_{entire\ field}$ computation.

center).[14] This idea relies on similar considerations to the one above: the surround is now composed of all three types of cell, and with the numbers of each cell type taken into account, that amounts to a weighted comparison of the center cell's type against the other two types. And just as in the previous case, the new and old computational structures are not the same. The new computations have three terms, making them as different from each other as $f(A)$ is from $f(A - B)$. And in the new computational structure, parts of the system that used to be performing the same M – S computation are now performing *two distinct* computations.[15]

In both cases, we have authors who emphatically accept that there were no color vision-relevant changes to the causal structure of the relevant circuits. In fact, these authors are precisely the ones I relied on above to clearly express that view. But in both cases the authors suggest a computational change in the post-mosaic circuits. And it's not a superficial change — it's not just that we have to call the cones something different because they're no longer 'L' cones. In the first case a spatially structured center vs surround computation became a non-spatially structured L vs M computation, with L and M representing weighted averages across the entire receptive field. And in the second case what used to be one computation is now two, and with extra (non-trivial) terms capturing different processes in post-mosaic circuits.

I've emphasized that externalism, in the sense I'm concerned with, is a claim about computation *in cognitive science*. And it seems that, as computation is understood in cognitive science, there is no assumption — in fact, there is outright rejection of the idea — that the computational structure of post-mosaic circuits in the pre- and post-mutation organisms must be the same, even when their causal structure is the same. So it would be a mistake to assume that ourselves, or to impose that assumption on this work.

The upshot of this subsection is that there is a case in neuroscience precisely mirroring the important aspects of the toy system. An organism undergoes a change in its causal structure at the extreme periphery, with no relevant changes aside from that. Nonetheless, this changes the computational structure of the organism, even in components whose causal structure is unchanged. What remains is to see whether the slight and peripheral difference in causal structure alone can ground the substantial difference in computational structure. If not, then the different computational structures would have to be grounded in features external to the organism itself, and externalism would be vindicated.

[14] These are really only half the possibilities. These cells can be center-ON surround-OFF *or* center-OFF surround-ON, and each entails a distinct pair of computations in the trichromat (Neitz and Neitz 2011, p. 646).

[15] Neitz and Neitz (2011) ultimately take an ecumenical view, accepting a role for what used to be shape perception circuits in addition to what used to be color perception ones: "red–green opponent signals in trichromats may be carried by two parallel pathways that preexisted in ancestral dichromatic primates, one for extracting contours relevant to spatial form and one that pre-existed for blue–yellow hue perception" (p. 648).

3.2 Internalism and externalism reconsidered

So far we've seen evidence that the computational structure of the circuits responsible for color vision depends on the make-up of the photoreceptor mosaic. This doesn't establish externalism directly, but it poses a challenge that internalists will have trouble meeting: to find a plausible basis for the computational difference between the pre- and post-mutation organisms — call them *Pre-M* and *Post-M*.

The externalist has a few options. We might think that an organism's computational structure is composed of *the aspects of its causal structure that explain the way it takes advantage of features of the environment*. Post-M can take advantage of differences between stimuli that Pre-M can't. If the organisms' computational structure must explain this difference, that entails a new computational structure for Post-M. More generally, the externalist is in a comfortable position here. She can appeal to facts about the environment and an organism's capacity to act in it and respond to it. And those are precisely the kind of things that computational structures are supposed to be responsible for, and to explain. That makes them natural, relevant, and well-motivated resources for understanding computational structure. But my purpose here is to sow pessimism about the *internalist's* ability to ground a computational difference between Pre-M and Post-M. (But see Richmond n.d.*a*, *b*, which develop an account of computational and representational explanation consistent with the externalism just described.)

The challenge for the internalist is a lack of plausible difference-makers corresponding to the different computational structures. The only relevant internal difference between Pre-M and Post-M is the new cone type in Post-M's retinal mosaics. But the difference-maker cannot simply be *the introduction of a third type of opsin or cone*. A new opsin could have a similar enough wavelength sensitivity profile to the existing ones that it would make no difference to color vision, or might differ from the existing ones merely in, e.g., the speed of its response — neither would call for computational re-description.[16] And even a new cone type with a significantly different wavelength sensitivity profile might not contribute to color vision — e.g., a cone sensitive to blue light that just feeds into a circuit for tracking day–night cycles (Horiguchi et al. 2012).

So the internalist's difference-maker has to be more complex. The strongest contender would combine all the specifics about the new retinal mosaic: there is a new type of cone, making three in total; the new cone type has a unique wavelength sensitivity profile; and the new cone type feeds into to the circuits that are responsible for color vision, just as the other cone types do. But even this isn't a plausible difference-maker. Many human females have not three but four or five

[16] 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.

types of cone cell that are connected to the color vision system, but this does not call for a novel computational structure (Jordan et al. 2010) — they appear to do the same L/M/S computations as the rest of us. Without going into the specific wiring details of Pre-M and Post-M's retinas, it's not clear what else the internalist can say. And the specific wiring details are no help either: the point of the previous subsection was that to the extent these detailed changed, they are irrelevant. The only differences in these wiring details between dichromats and trichromats are the changes that exist between *all* organisms, resulting from the random generation of those circuits. If Pre-M and Post-M have different computational structures for color vision simply because of the fine-grained differences in the wiring details of their retinas, then so would every other organism have a new and unique computational structure for color vision — a non-starter if we're trying to make sense of computations that are shared by trichromats but not dichromats (and vice versa).

The other options that might come to mind for the internalist fail more straightforwardly. Can the internalist take advantage of the *results* of the mutation, i.e., the fact that Post-M is trichromatic? That may be why computational re-description is required, but the standard definition of trichromacy refers to an agent's interactions with environmental light sources (recall the beginning of §3.1). So appealing to Post-M's trichromacy introduces environmental considerations: it is not a legitimate resource for the internalist. What about a different sort of difference between the organisms: not in their causal structure or "topology" (Chalmers 2011), but their *patterns of neural activation*. Certain cells in Post-M will be excited more or less often than in Pre-M, certain circuits' activity will be more or less correlated with others', and so on. That would count as an internal difference between the organisms, and it is consistent with there being no change to the post-mosaic circuitry. But this is subject to the same concerns we just saw. Not just any difference in patterns of neural activity makes a difference to computational structure. Reversing glasses, not to mention any temporary visual deprivation, cause significant changes to patterns of activity in the nervous system (Miyauchi et al. 2004; Harris 1965). But we don't think an organism's computational structure changes every time they close their eyes or put on a pair of reversing glasses. And besides, while patterns of neural activity are legitimate sources of *evidence* for computational structure, it is unclear that they are appropriate *grounds* for computational structure: a computer's patterns of internal activity change constantly depending on the inputs it is receiving, but its computational *structure* should be at least somewhat enduring (cf. Chalmers 2011; Rust and Movshon 2005; Yamins and DiCarlo 2016, also discussed below). In short, appealing to different patterns of activity between Pre-M and Post-M has all the problems that appealing to differences in their retinal mosaics did, plus some of its own.

To summarize, the computational difference between Pre-M and Post-M does not seem to be grounded by the resources available to the internalist, even when we specify the most clearly relevant properties of the new cone mosaic. That is not to say an internalist *couldn't* come up with

a way of grounding the computational difference between Pre-M and Post-M, but their limited resources, and the fact that the resources they do have don't seem up to the task, constitute a strong argument for externalism.

This is as good a point as any to return to an issue I postponed early on: my decision to interpret the difference between internalism and externalism as a disagreement about whether the grounds of a system's computational structure *refer to the environment*, rather than a disagreement about whether computational structure *supervenes on the system's internal states.* A supervenience externalist would have no reason to worry about Pre-M and Post-M. Because there is a difference in their internal make-up (however peripheral), their difference in computational structure is powerless to refute the idea that computational structure *supervenes* on a system's internal make-up. So, can the internalist revert to the supervenience definition of internalism and externalism?

Only, I think, at the cost of irrelevance. I tied externalism to a series of methodological problems in cognitive science about whether the environment is relevant to understanding the brain's computational structure. And regardless of those problems in particular, the goal is to understand computation as that notion is used in cognitive science. I don't think the supervenience definition can help us do that, for two reasons. To set them up, consider a pair of conflicting principles for determining a system's computational structure.

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

Broad Externalism. A system's computational structure is partly determined by features of its current environment, *and* by features of any other environment it might occupy (or any environment that is of interest to cognitive science).

Both refer to the environment, so they are externalist on the reference definition. But Broad Externalism 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 Broad Externalism doesn't privilege the current environment in any way (and let's stipulate that it doesn't), the computational structure of a system will not change as the system is moved from environment to environment. So the supervenience definition renders Broad Externalism *internalist*.

That might not be a problem if Broad Externalism was implausible, but it is in fact more plausible than Narrow Externalism. Compare the widely held desideratum that a theory of cognition should predict an organism's response to "arbitrary stimuli" (Rust and Movshon 2005; Yamins and

DiCarlo 2016), and not by inventing a new computational structure for each stimulus but with "a single theory that can predict neuronal and population responses to any arbitrary stimulus" (Rust and Movshon 2005, p. 1647). That is, if multiple environments are relevant they should determine a system's computational structure *throughout* those different environments — we should not have to create a new computational structure for each relevant environment. So definitions of internalism and externalism should be tested by what they say about Broad Externalism, not (just) Narrow Externalism.

So the first problem with the supervenience definition is that if Broad Externalism or something like it is operative, the supervenience definition lets us get out of the externalism debate — the debate over whether the environment helps determine the brain's causal structure — without having to decide whether the environment helps determine the brain's causal structure. Either it doesn't, in which case internalism is true. Or it does, in which case internalism is true again because (as I described two paragraphs up) the supervenience definition counts Broad Externalism as internalist. The *reference to the environment* definition defines internalism and externalism in a way that carves the debate more cleanly: a view is externalist just as long as it gives the environment some place in the factors that determine a system's computational structure. On the reference definition, to establish externalism we have to engage with the central question that the supervenience definition could bypass: what factors help determine a system's computational structure?

The second problem with the supervenience definition is that it pulls us impossibly far away from the methodological debates I was aiming to intervene on, which asked whether the environment is relevant for discovering and understanding a system's computational structure. What guidance can philosophical externalism offer? The supervenience definition would tell cognitive scientists to run around a bunch of different environments and see whether their target system's computational structure changed. But they wouldn't know whether the computational structure had changed unless they had a way of knowing what that computational structure was in the first place. And *that* would already be an answer to their methodological questions. So the supervenience definition can't help answer neuroscience's methodological questions — it can't even be relevant to neuroscience until those questions are answered. And if those methodological questions were answered, and neuroscientists knew how to determine a system's computational structure, what would the supervenience definition add? Internalism or externalism on that definition would be a claim about how a system's computational structure in one environment is or isn't different than its computational structure in another environment. But if the methodological question is answered, and neuroscientists know how to determine a system's computational structure, they will already have this information. They will already be able to determine a system's computational structure in the one environment, and then in the other, and then compare the two. The supervenience definition

of externalism is therefore not an attractive option for the internalist unless they want to set aside the methodological debates I've targeted, and likely any interesting connection to computation as that notion is used in cognitive science.

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 externalism along those lines would have to be derived from the way the environment supports and constrains computational explanations, and there is much work left to do on that front. My argument has simply been against internalism, and for externalism *simpliciter*. But even the bare thesis of externalism provides a way into the methodological debates I began with, and I'll return to those debates now.

4 Upshots, methodological and otherwise

If externalism is wrong and the brain has its computational properties intrinsically, then we should be able to discern those properties by studying the brain itself (contra Krakauer et al. 2017). We could potentially do that with connectomics, and without worrying about naturalistic stimuli except insofar as they drive the brain in ways that illuminate more of its internal properties. But if externalism is right and the brain has its computational structure only extrinsically, then this strategy is not plausible. We should not expect to glean the brain's computational properties by studying the brain alone. The processes governing behavior are not best inferred from examination of the processors. They *cannot* be inferred from examination of the processors. Any inference to an organism's computational processes or structure needs to consider the parts of the environment they depend on.

Likewise, if externalism is right, even the extreme detail connectomics promises will not reveal the brain's computational structure. 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. But the appreciation of connectomics is misguided insofar as it derives from a belief that the connectome will settle questions about the computational structure of the brain, or will settle those questions without a careful understanding of the brain's environment.

This point 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 from the also common (and also well-taken) criticism that connectome maps don't indicate the strength of connections, whether synapses are inhibitory or excitatory, and so on (Morgan and Lichtman 2013). The point is that even if we had all those details, connectomics alone would be unable to tell us the brain's computational structure. It is often pointed out that even the highly detailed connectome maps of C. elegans provide scant understanding of its behavior (Niv

2020, 134), and externalism explains why: assuming an organism's behavior is to be explained by its computational structure, that structure just ain't in the connectome.

These are already significant methodological upshots. They limit and clarify the explanatory power of connectomics, and direct attention to the environment rather than the brain alone. In particular, they weigh in on the side of Krakauer et al. (2017) and Niv (2020): neuroscience needs a more thorough understanding of the environment and an organism's behavior in it, not just increasingly detailed descriptions of the organism's nervous system.

The upshots regarding naturalistic stimuli are looser, but still important. If externalism is right, our choice of stimuli should depend on more than their ability to reveal the brain's intrinsic properties. A stimulus will have features that do or don't help determine a brain's computational structure, and knowing which features are which would tell us which stimuli have a deep relationship to the brain's computational structure, and which merely drive neural activity. A more fleshed-out version of externalism would tell us which aspects of the environment, and of potential stimuli, help determine the brain's computational structure, and which don't. But even the bare thesis of externalism implies that our choice of experimental stimuli must reflect our understanding of how the relationship between brain and environment grounds the brain's computational structure.

An example from color vision will help flesh out externalism's role in all these methodological debates. Recent technology 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 useful for understanding physiological details and causal structure in the retina, but it is also taken to reveal the *computational* structure of the early visual system (Kling et al. 2019). If externalism is true, that kind of inference from causal and physiological structure to computational structure is undermined for the same reason the inference from connectomics to computational structure is undermined. That inference has to consider the environmental, not just internal, determinants of computational structure. And a weaker but still important point applies to the stimuli used. Single-cone stimulation is highly unnatural: because of optical blur imposed by the cornea, no scene outside the laboratory is ever represented in the retina at single-cone resolution (Kling et al. 2019). The further development of externalism is necessary to understand *how* and *why* this kind of non-naturalistic stimulus can be informative about the brain's computational structure, beyond its ability to reveal physiological and computational structure by driving neural activity. So externalism *simpliciter* rules out some inferences concerning computational structure that neuroscientists have been tempted to make, and the further development of externalism is necessary to reveal the true significance of single-cone stimulation experiments and their stimuli.

All the same upshots hold for explainable AI, insofar as it is able to use a notion of computation similar to neuroscience's to help make the behavior of neural networks intelligible. Computational

structure in this sense will not be a matter of a network alone, but also of its environment. That would explain why neural networks are often uninterpretable despite our perfect access to their nodes and weights, or their "connectome:" assuming their behavior will be explained by their computational structure (again, in the neuroscientific sense of that notion), computational structure isn't a matter of the connectome alone. This means that to understand AI systems we must carefully study their environments, e.g. their training data, test data, and the data in the environments they will be deployed in. Understanding that data will not just help us predict how a model's performance will generalize (Herrmann et al. 2024; Strobl and Leisch 2022; Koch et al. 2021). Understanding that data will also be necessary to understand the structures in the model that *generate* its performance. The data will be part of what determines which aspects of the model's fine-grained structure rise to the level of *computation* in the neuroscientific sense, supporting and explaining the aspects of behavior that computational structure paradigmatically does.

Externalism also means we need to reject inferences to computational structure that are based solely on the details of an AI system itself. E.g., we could not conclude that a large language model uses a *world model* based solely on an examination of its vector representations (cf. Yildirim and Paul 2024). To count as part of its computational structure, that world model would have to be related in appropriate ways to the model's tasks and the environment it performs them in. Concerning the relevance of different sorts of data or environments (analogous to "naturalistic" stimuli and environments), consider Morrison's suggestion that a model's computational structure depends partly on the tasks it performs on its current data, and partly on how quickly it *adapts* to perform related tasks with different data (Morrison n.d.). In that case it will be essential to decide which tasks and datasets help determine a network's computational structure — on which tasks and given which datasets does the model's adaptation reveal something about its computational structure? The answer might appeal to the environments with respect to which we're interested in explaining the model's behavior, or the environments or features of the environment that are especially relevant to stakeholders in the network's behavior. In either case, it will be necessary to develop externalism — to say exactly how a model's computational structure depends on which aspects of its environment — in order to grasp its computational structure.

There is much more to say about the relationship between externalism and scientific practice. The point in this final section has been that externalism, properly understood, answers some important questions in neuroscience and provides a starting point to answer others. The upshots for philosophy are also significant. We have untied two definitions of internalism and externalism, and seen how the *reference to the environment* definition makes the debate over externalism informative about and useful to cognitive science. We have reinforced an interpretation of externalism as a view about the *determination* rather than the *classification* of computational structures, to similar effect.

We have taken the argument for structural externalism beyond toy examples and into the details of scientific practice. And we have seen that representationalism about computation is not necessary to establish externalism, and nor is a syntactic or causal view of computation sufficient to establish internalism (as I've only assumed that computational structure is some type of causal structure).

In fact, we haven't just gotten by without representationalism. We've gotten by without *any* view concerning the nature or definition of computation — representational or teleological or mechanistic or otherwise. These debates are central to the philosophy of computation (Piccinini 2015; Shagrir 2022; Anderson and Piccinini 2024). But they were not necessary to generate the list of implications above. Instead, we have looked in detail at the investigations and research programs that use a notion of computation, without attempting to define that notion or give an account of a *kind* it might refer to. This may be another important upshot, in line with a deflationist approach to computation (Richmond n.d.*a*; Curtis-Trudel 2024; Williams 2024) and to cognitive science generally (Egan 2021; Richmond n.d.*b*; Cao 2022). We can make significant progress in understanding externalism, and computational explanation more broadly, without getting tied up in questions about *what it is* to compute, and by focusing instead on how the notion of computation supports the scientific investigations that use it: what it helps scientists do, and how.

References

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

Anderson, N. G. and Piccinini, G. (2024), *The Physical Signature of Computation*, Oxford University PressOxford.

Bolt, T., Anderson, M. L. and Uddin, L. Q. (2018), 'Beyond the evoked/intrinsic neural process dichotomy', *Network Neuroscience* **2**(1), 1–22.

Bontly, T. (1998), 'Individualism and the Nature of Syntactic States', *The British Journal for the Philosophy of Science* **49**(4), 557–574.

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

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

Chalmers, D. J. (2011), 'A Computational Foundation for the Study of Cognition', *Journal of Cognitive Science* **12**(4), 323–357.

Chang, L., Breuninger, T. and Euler, T. (2013), 'Chromatic Coding from Cone-type Unselective Circuits in the Mouse Retina', *Neuron* **77**(3), 559–571.
URL: *http://dx.doi.org/10.1016/j.neuron.2012.12.012*

Chemero, A. (2009), *Radical Embodied Cognitive Science*, MIT Press.

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

Cornelissen, F. W. and 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.

Crick, F. H. (1979), 'Thinking about the brain.', *Scientific American* **241**(3), 219–232.

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

Curtis-Trudel, A. (2024), 'Computation in Context', *Erkenntnis* .

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A., Kriegeskorte, N. and Kietzmann, T. C. (2023), 'The neuroconnectionist research programme', *Nature Reviews Neuroscience* **24**(7), 431–450.

Dulai, K. S., Von Dornum, M., Mollon, J. D. and 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.

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

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

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

Elam, J. S., Glasser, M. F., Harms, M. P., Sotiropoulos, S. N., Andersson, J. L., Burgess, G. C., Curtiss, S. W., Oostenveld, R., Larson-Prior, L. J., Schoffelen, J. M., Hodge, M. R., Cler, E. A., Marcus, D. M., Barch, D. M., Yacoub, E., Smith, S. M., Ugurbil, K. and Van Essen, D. C. (2021), 'The Human Connectome Project: A retrospective', *NeuroImage* **244**.

Feldmann, M., Beckmann, D., Eysel, U. T. and 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* pp. 1–14.

Fletcher, S. C. (2018), 'Computers in Abstraction/Representation Theory', *Minds and Machines* . **URL:** *https://doi.org/10.1007/s11023-018-9470-9*

Gould, S. J. and 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.

Hardin, C. L. (1988), *Color for Philosophers*, Hackett.

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

Herrmann, M., Julian Lange, F. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., Boulesteix, A.-L. and Bischl, B. (2024), Position: Why We Must Rethink Empirical Research in Machine Learning, *in* 'Proceedings of the 41st International Conference on Machine Learning'.

Horiguchi, H., Winawer, J., Dougherty, R. F. and 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. and 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. and Niell, C. M. (2011), 'What can mice tell us about how vision works?', *Trends in Neurosciences* **34**(9), 464–473.
URL: *http://dx.doi.org/10.1016/j.tins.2011.07.002*

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. (2018), 'Photopigments and the dimensionality of animal color vision', *Neuroscience and Biobehavioral Reviews* **86**, 108–130.
URL: *https://doi.org/10.1016/j.neubiorev.2017.12.006*

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

Jacobs, G. H., Williams, G. A., Cahill, H. and Nathans, J. (2007), 'Emergence of Novel Color Vision in Mice Engineered to Express a Human Cone', *Science* **315**(March), 1723–1725.

Jonas, E. and Kording, K. P. (2017), 'Could a Neuroscientist Understand a Microprocessor?', *PLoS Computational Biology* **13**(1), 1–24.

Jordan, G., Deeb, S. S., Bosten, J. M. and Mollon, J. D. (2010), 'The dimensionality of color vision in carriers of anomalous trichromacy', *Journal of Vision* **10**(8), 12.
URL: *http://jov.arvojournals.org/Article.aspx?doi=10.1167/10.8.12*

Jordan, G. and Mollon, J. (2019), 'Tetrachromacy: the mysterious case of extra-ordinary color vision', *Current Opinion in Behavioral Sciences* **30**, 130–134.
URL: *https://doi.org/10.1016/j.cobeha.2019.08.002*

Kelber, A. (2016), 'Colour in the eye of the beholder: receptor sensitivities and neural circuits underlying colour opponency and colour perception', *Current Opinion in Neurobiology* **41**, 106–112.
URL: *http://dx.doi.org/10.1016/j.conb.2016.09.007*

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

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

Koch, B., Denton, E., Hanna, A. and Foster, J. G. (2021), Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research, *in* '35th Conference on Neural Information Processing Systems'.
URL: *https://paperswithcode.com*

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

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

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

Mancuso, K., Hauswirth, W. W., Li, Q., Connor, T. B., Kuchenbecker, J. A., Mauck, M. C. and Neitz, J. (2009), 'Gene therapy for red-green colour blindness in adult primates', *Nature* **461**, 784–788.

Mancuso, K., Mauck, M. C., Kuchenbecker, J. A., Neitz, M. and Neitz, J. (2010), 'A Multi-Stage Color Model Revisited: Implications for a Gene Therapy Cure for Red-Green Colorblindness', *Advances in Experimental Medicine and Biology* **664**, 631–638.

Markram, H. (2006), 'The blue brain project', *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC'06* **7**(February), 153–160.

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

Merel, J., Aldarondo, D., Marshall, J., Tassa, Y., Wayne, G. and Ölveczky, B. (2019), 'Deep neuroethology of a virtual rodent', *Draft* pp. 1–20.
URL: *http://arxiv.org/abs/1911.09451*

Miyauchi, S., Egusa, H., Amagase, M., Sekiyama, K., Imaruoka, T. and Tashiro, T. (2004), 'Adaptation to left – right reversed vision rapidly activates ipsilateral visual cortex in humans', *Journal of Physiology* **98**, 207–219.

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. and Lichtman, J. W. (2013), 'Why not connectomics?', *Nature Methods* **10**(6), 494–500.

Morrison, J. (n.d.), 'Rules for Nodes and Neurons', *in draft* .

Naddaf, M. (2023), 'Europe spent €600 million to recreate the human brain in a computer. How did it go?', *Nature* **620**(7975), 718–720.

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

Neitz, J. and Neitz, M. (2011), 'The genetics of normal and defective color vision', *Vision Research* **51**(7), 633–651.
URL: *http://dx.doi.org/10.1016/j.visres.2010.12.002*

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

Niv, Y. (2020), On the Primacy of Behavioral Research for Understanding the Brain, *in* A. J. Lerner, S. Cullen and S.-J. Leslie, eds, 'Current Controversies in Philosophy of Cognitive Science', Routledge, pp. 134–149.

Peacocke, C. (1992), *A Study of Concepts*, MIT 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. and Shagrir, O. (2014), 'Foundations of computational neuroscience', *Current Opinion in Neurobiology* **25**, 25–30.
URL: *http://dx.doi.org/10.1016/j.conb.2013.10.005*

Putnam, H. (1975), 'The Meaning of "Meaning"', *Minnesota Studies in the Philosophy of Science* **7**, 131–193.

Pylyshyn, Z. W. (1993), Computing in Cognitive Science, *in* M. I. Posner, ed., 'Foundations of Cognitive Science', MIT Press, pp. 49–92.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. Roberts, M. E., Shariff, A., Tenenbaum, J. B. and Wellman, M. (2019), 'Machine behaviour', *Nature* **568**(7753), 477–486.
URL: *http://dx.doi.org/10.1038/s41586-019-1138-y*

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

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

Richmond, A. (n.d.*b*), 'What is a theory of neural representation for?', *Synthese (forthcoming)* .

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

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

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

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

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

Shagrir, O. (2020), 'In defense of the semantic view of computation', *Synthese* **197**, 4083–4108.
URL: *https://doi.org/10.1007/s11229-018-01921-z*

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

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

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

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

Sonkusare, S., Breakspear, M. and Guo, C. (2019), 'Naturalistic Stimuli in Neuroscience: Critically Acclaimed', *Trends in Cognitive Sciences* **23**(8), 699–714.
URL: *https://doi.org/10.1016/j.tics.2019.05.004*

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

Strobl, C. and Leisch, F. (2022), 'Against the "one method fits all data sets" philosophy for comparison studies in methodological research', *Biometrical Journal* **66**(1).

Wachtler, T., Dohrmann, U. and Hertel, R. (2004), 'Modeling color percepts of dichromats', *Vision Research* **44**, 2843–2855.

Wachtler, T. and Wehrhahn, C. (2016), Computational Modeling of Color Vision, *in* J. Kremers, R. C. Baraas and N. J. Marshall, eds, 'Human Color Vision', Springer, pp. 243–268.

Wallisch, P. and Movshon, J. A. (2008), 'Structure and Function Come Unglued in the Visual Cortex', *Neuron* **60**(2), 195–197.
URL: *http://dx.doi.org/10.1016/j.neuron.2008.10.008*

Williams, D. (2024), 'It takes two to make a view go right'.
URL: *https://philosophyofbrains.com/2024/10/03/10-13140-rg-2-2-35310-63041.aspx*

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

Yildirim, I. and Paul, L. A. (2024), 'From task structures to world models: what do LLMs know?', *Trends in Cognitive Sciences* **28**(5), 404–415.

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