**Computational Externalism**

Andrew Richmond

**Abstract**

Many cognitive systems compute—at least, we explain them and their capacities by appealing to the computations they implement. For those explanations to work, they must appeal to computations those systems actually implement, and to know, in a general way, which computations systems implement, we will have to know which states of affairs make it the case that a system implements the computations it does; which properties we must attend to in order to discover the system’s computational identity.[[1]](#footnote-1) One aspect of this question is the internalism–externalism question: *where* are the states of affairs are that determine a system’s computational identity? Internalists argue that states of affairs internal to a system fullydetermine its computational identity; externalists argue that states of affairs external to the system must also contribute. I pose a challenge for internalism, derived from a case study in the theory of trichromacy, and I argue that internalism won’t likely meet the challenge. Along the way I argue for definitions of internalism and externalism more fruitful than the ones in common usage, criticize some trends in the philosophy of computation, sketch some implications of externalism for cognitive scientific methodology, and prepare the way for a scientifically informed elaboration of computational externalism.

**1 Introduction**

Computational explanationis ubiquitous in the cognitive sciences. A computational explanation takes as explanandum some capacity of a cognitive system, and offers as explanans some computation performed by that system. The system—usually a brain or nervous system—is supposed not only to be *modelled* by the computation, but to actually perform or *implement* it. That means we need to know which computations a system implements in order to know which computational explanations of it are correct. An essential first step in a theory of implementation is to decide what kinds of properties determine the computations that a system implements. In particular, we want to know *where* the properties are that determine a system’s computational identity. Are some of those properties external to the system—in its environment, e.g.—or are only properties of the system itself relevant? The former position is *externalism*, the latter is *internalism*.

 The purpose of this article is two-fold. First, and most explicitly, I will make a new argument for computational externalism, one that places that thesis in tight connection with the practice of computational explanation in the cognitive sciences. But I also intend to be making a fresh start on computational explanation. As will become clear, I set aside many of the assumptions the current literature makes—few of which seem destined for consensus, and most of which incur significant explanatory debts themselves. Instead, my only assumption about the notion of computation—as that notion is used in cognitive science—will be that it must support the kinds of judgments, inferences, and classifications that are actually made as cognitive scientists give and justify their explanations. Applying this minimal assumption to a case study will furnish the argument for externalism, and in the course of the article that assumption will also lead to an unorthodox definition of externalism, clarification of the notion of computation itself, and a number of observations about methodological issues surrounding computational explanation and the way philosophy can intervene in them.

I begin by outlining the notion of computation and its significance for the cognitive sciences (section 2). I then clarify the notion of externalism, with special attention to its relevance for methodological questions in the cognitive sciences (section 3). Then I move on to a case study of some results in the computational theory and evolutionary biology of color vision (section 4). I argue that those results pose a challenge that internalism is unlikely to meet (section 5), and conclude by considering the impact of externalism on cognitive scientific methodology more broadly (section 6).

**2 Computation and Computational Explanation**

Many disciplines use computational models, but in the cognitive sciences the models are often understood to be theories about what computations the system being modelled is actually performing.[[2]](#footnote-2) That is, we claim that cognitive systems *implement* certain computations, and that this fact explains some of their capacities. We say that one system behaves the way it does because it computes sums, another because it computes the disparity between retinal images, and so on.

There are many theories of implementation—of what a system must be or do in order to implement a computation. At the very least, it seems necessary that the system go through a process that mirrors the stages of the proposed computation (e.g. Cummins, 1991; Egan, 2010, 2014; Pylyshyn, 1993, p. ??). If I say some system is adding pairs of numbers, the process it undergoes had better have parts (states or components) that can be identified as the addend1-part, the addend2-part, and the sum-part, and it must reliably transition from the addend-states to the sum-state, or use the states of the addend-components to determine the state of the sum-component. Otherwise, even if the system’s outputs can be predicted by the computation, the computation isn’t going on inside the system. This *minimal condition* is all we’ll need to generate the externalist result below, so I won’t assume that computing systems have any features other than these—they have parts or components that map to the algorithms the system implements.

 Working with just the minimal condition also enables me to make two general points. First, most of the externalist literature appeals to notions of representation or function (Fletcher, 2018; e.g. Peacocke, 1994, 1999; Piccinini, 2008, 2015; Piccinini & Shagrir, 2014; Rescorla, 2013; Shagrir, 2001, 2018; Shea, 2013; Sprevak, 2010), but we’ll see that externalism can be defended without appealing to those notions, and doesn’t depend for its plausibility on theirs. In fact, externalism as defended in this paper provides a test for more committal theories: they should preserve the independently derived externalist result that I discuss, along with others of the same ilk. And second, contrary to the assumptions of the internalist literature (see esp. Egan, 1991, 1995, 1999, 2010, 2014), a formal or mapping account of computation does not suffice to establish internalism.[[3]](#footnote-3)

 Using just the minimal condition also makes the conclusion of this paper more surprising. There are two sorts of computational externalism. One is employed by theorists who, like some of those cited above, think computations must be performed on entities that represent external things. If representation underwrites an essential aspect of computation’s explanatory role, then it is natural to say two computations differ if their components represent different external things. This would make you an externalist.

But there is another thesis, *syntactic externalism*, much closer to the one I’ll be defending. The syntactic structure of a system is what you would get if you took a computation over representational entities and stripped away the representational content. You’d be left with computational *vehicles*, or bearers of content: the states or components of a system that correspond to the algorithms it implements (see the minimal condition above). The syntactic, or vehicular, externalist says that a computational system’s structure, even at this level, is externally determined. There is an overwhelming consensus in the literature that this cannot be the case, that a system’s vehicular structure can only depend on features of the system itself, particularly the structure and relations of its computational vehicles, which are generally supposed to be physical parts (Shea, 2013). This is natural enough, given that those vehicles connect a system’s computations to its causal structure, which underwrites the stability and predictability of the computations, not to mention their objectivity (Shea, 2018, pp. 37–38). That causal structure is, intuitively at least, a state of affairs inside the system.

But as I said, I won’t even assume that computations *have* features other than their formal or syntactic structure—their being carved up into parts that map to a certain algorithm. So by default it is that structure that I’ll be arguing is determined externally. Of the few theorists who have accepted this claim—as far as I know, just Thomas Bontly (1998), Amir Horowitz (2007), Oron Shagrir (2001, 2018), and Nicholas Shea (2013)—all but Bontly argue for it on the basis of a representational account of computation (and Bontly invokes teleological considerations). It would be more convincing if we had a non-representational (and non-teleological) justification of the view—one that doesn’t hinge on the notion of representation, or on representational content being of external rather than internal things, and so on. It will also be important that my argument ties externalism closely to the actual practice of computational explanation in cognitive science, and some examples of it—something sorely missing in the above accounts.

Before moving on, let me make a caveat and a give couple definitions. The caveat is this: I’m concerned with computational explanation in actual scientific practice. A move that says, “Cognitive scientists use the notion of computation like A, but actually computation is B” has a huge hurdle to clear: it has to show either that what cognitive scientists really mean when they talk about computation is B, or that B does a better job explaining cognitive science’s explananda than A does. Where neither of those conditions is met, the move is illegitimate. Where I need to refer to this mistake later, I’ll call it the bad move. (“The But-actually-computation-is-B Fallacy” doesn’t quite trip off the tongue.)

 Here are the definitions. By *computation* I mean algorithm (as in Marr Level 2, not Marr Level 1 (Marr, 1982)). Algorithms are sequences of transitions between abstract states. This is a mathematical notion. No physical system *is* an algorithm; nor is any part or process of it. Physical systems perform,or *implement*, algorithms. Your calculator is notan algorithm; it *implements* an algorithm (or many algorithms). The question at hand is what makes it the case that it implements the algorithms it does.

I’ll often use the terms *computational structure* and *computational identity*. Your calculator may currently be in the process of adding 5 to 7, but computational theories explain not just *instances* of computation—like summing two specific numbers, or performing the algorithm defined by the formula 5 + 7—but the more general capacity that underlies that summing: the ability to perform addition, to sum two arbitrary (or relatively arbitrary) numbers. This is what I mean by computational *structure* or *identity*, as opposed to an instance of computation.[[4]](#footnote-4)

To conclude this section, let me say something positive about vehicular externalism—about how externalism even *could* be true, since it is highly unintuitive to most philosophers. There are different ways to articulate syntactic externalism, but all of them work by pulling the computational identity of a system away from the fine details of its causal structure. One could suppose, with Shagrir (2001), that a computational explanation simplifies a system’s structure by *collapsing* different physical states into equivalence classes. So a circuit with possible inputs of 0, 1, and 2 volts can be treated as though it has three kinds of input—0v, 1v, and 2v—or as if it has just two kinds of input—either 0-1v and 2v, or 0v and 1-2v. Certain circuits look like an AND-gate under the one grouping and an XOR-gate under the other (Shagrir, 2001). It isn’t clear whether or not this correctly describes any cognitive system (as opposed to the toy examples Shagrir uses), but it is possible for different groupings of states to be equally viable descriptions of a cognitive system’s vehicular structure, requiring external criteria to decide between them.[[5]](#footnote-5) Then the system’s vehicular structure—e.g. whether 0v and 1v pulses are the same vehicle, the same kind of input/state/component, or not—depends on features of the world external to the system.

Another option is to say that computational explanations simplify the structure of a system by *ignoring* certain of its operations in order to highlight others. Computational explanations then work by picking out important parts of a system, or parts that are particularly relevant to some phenomenon, and which parts are important or relevant—and count as vehicles—can’t be decided without reference to the environment. A more radical suggestion is that computational explanations outright falsify the structure of the system in order to latch on to important regularities or other features of the system’s information-processing structure or capacities, and idealizations are not chosen just on the basis of internal structure. I won’t argue for any of these here—at this point I just want the rough shape of vehicular externalism to be relatively clear.

**3 Internalism and Externalism**

A theory of implementation would tell us under what circumstances any system has any particular computational structure. As I said, my question here is narrower: *where* are the propertiesthat determine a system’s computational structure? We can pose the question differently. 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? This is a question of taxonomy or individuation, 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 of the world that *classify* the system as performing the computation it does. If those features can include ones external to the system, externalism is true. If they cannot, internalism is true.[[6]](#footnote-6)

 There are different ways to operationalize the notions of internalism and externalism. See this passage from Frances 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. (1994, p. 258, emphasis mine)

The two definitions, S and R, are generally taken to be equivalent. When they are not explicitly taken to be equivalent, S is usually preferred or assumed (e.g. Chomsky, 2000; Egan, 1992, 2003, 2010; Fodor, 1981; Stich, 2010) [Sterelny 1990, p. 45].[[7]](#footnote-7) I think this is a mistake. Not only are the definitions distinct; R is a much more useful notion for understanding and advancing 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. Imagine a taxonomic principle that said something like the following:

T1 A system’s computational structure is determined by certain relations it has to its environment, and the relations it *would* have to its environment in all possible environments.[[8]](#footnote-8)

T1 makes essential reference to the environment, so it is externalist by the standards of R. But it also ensures supervenience on internal properties: every environment that the system can be moved to or imagined in is already taken into account in the computational identity of the system in its current environment. As long as the principle doesn’t privilege the current or actual environment in any way (and lets stipulate that it doesn’t), the computational structure of a system will not change from environment to environment: the principle is internalist by the standards of S.[[9]](#footnote-9)

 So the two definitions come apart. Leaving aside the possibility of a third, which of S and R should we prefer? As I mentioned, my concern with computation is with how the notion is and can be used in scientific practice. So we should look to situations where internalism and externalism are relevant to that practice. The clearest example is the perennial debate about the relevance of ethology to neuroscience, recently revitalized by John Krakauer and colleagues (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). The question is whether we can understand a cognitive system if we neglect to carefully consider its environment and its relationship to that environment. For Krakauer et al (and most cognitive scientists) this is a methodological issue, rather than a taxonomic or individuative one. Should we worry about designing ecologically valid task-environments? Must we do detailed task analysis to understand the results of neuroscientific experiments?[[10]](#footnote-10) But the individuative question bears on the methodological one: if externalism is true, we had better act—experiment, theorize—like it is. When we design and analyze experiments to investigate the computational structure of a cognitive system, we had better worry about the parts of the environment that are relevant to determining that computational structure.[[11]](#footnote-11)

 So which definition of externalism—S or R—promises to shed light on the issues raised by Krakauer et al? If T1 or something like it is operative, S is of little help: internalism is guaranteed, regardless of the role the environment might play in determining the nature of the system we’re studying. That secures internalism (or ‘internalisms’) but prevents us from addressing the questions at hand: how the environment might contribute to the phenomena under investigation; what parts of the environment we should be paying attention to in experimental design and analysis; etc.

This is not just logic-chopping or definition-mongering. T1 is extreme, but compare, e.g., the widely held desire that a theory of cognition predict an organism’s response to “any arbitrary stimulus” (Rust & Movshon, 2005, p. 1647).[[12]](#footnote-12) Cognitive scientists don’t think about every environment or all possible stimuli all the time. Sometimes they test the perception of shapes with constant color; sometimes they test the perception of colors with constant shape. But they need to make sense of an organism’s behaviour across both environments, not to mention ones where color and shape both vary. We don’t want organisms to have different computational structures in all the different environments they encounter, or for all the different tasks they perform.[[13]](#footnote-13) A unified understanding of behaviour across environments—maybe not all, but many—is at least a long-term goal of computational theorizing. So if the environment *is* involved in computational individuation, a plausible taxonomic principle is a restricted version of T1:

T2 A system’s computational structure is determined by certain relations it has to its environment, and the relations it would have in certain non-actual environments.

This is not yet to say that T2 is correct, only that it is a plausible version of the externalist position. And S renders T2—the externalist hypothesis—*internalist*, because the considerations determining a system’s current computational structure already take into account the environments it might be moved to or imagined in, meaning that when it is so moved or so imagined, its computational structure does not change. In short, S allows us to be ‘internalists’ about computational identity only by keeping us blind to the environment’s contribution to that identity.[[14]](#footnote-14) The real debate—over the details of T2, and over the questions raised by Krakauer et al—is bypassed. R, on the other hand, counts T2 as externalist, focusing the debate on precisely the questions that S is blind to, and making room to ask which features of the environment a taxonomic principle must make reference to, which features must be considered in experimental design, and so on. So when I come to the argument for externalism in section 5, and when I go on to consider implications for cognitive scientific methodology, I’ll be considering R—a version of externalism that says our taxonomic principles, or the basis on which we class systems as having the computational structure they do, make reference to the systems’ environments.

**4 Case Study: Color Vision**

So much for the preliminaries. This section discusses a finding in the science of color vision. The finding should be surprising, though it won’t seem to support externalism on its own—the rest of the camel comes in section 5. The finding is this: *the evolution from dichromacy to trichromacy occurred without any changes to post-retinal circuitry, but it calls for changes to the computational classification of certain 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 the section by considering some philosophical and empirical objections to the finding.

Humans, along with the rest of the old world primates, are behavioural trichromats, meaning that 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 are required, but for some all three are needed.[[15]](#footnote-15) (The definition of dichromacy simply replaces “three” with “two”.)

There are some common misconceptions about this definition. 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, Deeb, Bosten, & Mollon, 2010; M. Neitz & Neitz, 2014). Second, it is often assumed that the three dimensions mentioned in the definition of trichromacy are hue, saturation, and lightness. In fact, they are not. We can distinguish hue alone on more than one dimension (see Derrington, Krauskopf, & Lennie, 1984 for one illustration of this), and dichromat humans do not lack the ability to detect differences in hue, saturation, or lightness (Wachtler, Dohrmann, & Hertel, 2004). (The relation of the dimensions involved in trichromacy to the qualitative dimensions of hue, saturation, and lightness is still something of a mystery (MacLeod, 2010; Wachtler et al., 2004).)

The standard approach to color vision is summarized by Almut Kelber 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. (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* that they express. Opsins are molecules that react to light, triggering a change in the membrane potential of their cell, which causes it to send signals to post-receptoral 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, in humans and our closest relatives, circuits that perform some simple computations. One is

*a*L(λ) – *b*M(λ),

where L(λ) and M(λ) are the activity of L and M cones for a particular wavelength or wavelength profile, λ, and *a* and *b* are weightings of the L and M cone responses (Shevell & Martin, 2017). This is a computation of a weighted difference between L and M photoreceptor activation for a specific incoming wavelength or wavelength profile. 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 (e.g. 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. In primates the computations look something like the above, so I’ll use those computations for illustrative purposes. But everything I say can be translated to different models.

 These computations explain a lot about human color vision. They explain behavioural data to do with unique hues, color mixing, and opponent colors, among other phenomena (for two classical examples, see Devalois & Devalois, 1993; Hurvich & Jameson, 1957). To give just one example in more detail, the L – M computation is also known 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 apparently 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 largely independent of the red–green one, so 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.[[16]](#footnote-16)

 The computational model above is also supported by physiological findings. It generates 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 (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, Eskew, Martin, & Stockman, 2018; Jacobs, 2014). Another ongoing project investigates the weights of L, M, and S cell input to 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, Breuninger, & Euler, 2013; Sabesan, Schmidt, Tuten, & Roorda, 2016). The particulars of these last two paragraphs are not essential to the following; they are intended as a brief illustration of the depth and interest of this area of research, none of which would be possible without the computational model above. The computations that model ascribes to the brain (especially retinal and LGN circuits) are simply indispensable in the theory and investigation of color vision.

 But evolutionary biology raises important questions about these computations. We need a little more background first. In primates and other mammals the S opsin is coded for on a non-sex chromosome, which every member of the species receives. The M and L opsins are coded for on the X chromosome, so males of the species get just one copy of those, while females get two. Our dichromat ancestors had only one of the longer-wavelength opsins, the M opsin.[[17]](#footnote-17) 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. Whichever X chromosome a female’s cone cells expressed (due to random X-inactivation, about half express one chromosome and half express the other (Jacobs, 2008)), the result is a retina that looks just like the males’.

 The surprising finding in this literature is that the evolution of trichromacy from dichromacy seems to have occurred *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 an allele—the L opsin gene (Dulai, Von Dornum, Mollon, & Hunt, 1999; Jacobs, 2009; Jacobs & Nathans, 2009). That is, trichromacy evolved from dichromacy immediately on a change to the photoreceptor mosaic, without any change to post-retinal circuitry.[[18]](#footnote-18) This is widely endorsed in the literature,[[19]](#footnote-19) and it matters for our purposes because *the computations we ascribe to the dichromat and the 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.[[20]](#footnote-20) 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; it’s that 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 in a moment, but let me first set out the evidence for the finding that trichromacy developed immediately on the introduction of L cones to the retina. Evidence from the similarity of the M and L opsin genes, along with their locations on the X chromosome, strongly suggests that—as I said above—one resulted from a transcription error copying the other (Dulai et al., 1999; Jacobs, 2009; Jacobs & Nathans, 2009). It likely 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 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 have immediately resulted in a new dimension of color vision[[21]](#footnote-21) (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, Mauck, Kuchenbecker, Neitz, & Neitz, 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). 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 squirrel monkeys. 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 in knock-out mice, and at adulthood heterozygote females—with both alleles, and thus a three-opsin mosaic—showed an extra dimension of color vision, discriminating between colors that their two-opsin conspecifics could not (Jacobs, Williams, Cahill, & Nathans, 2007).

 The second experiment was performed on 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.[[22]](#footnote-22) Male squirrel monkeys are therefore dichromats, and females who are lucky enough to 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 and 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. And as I said above, the move from dichromacy to trichromacy calls for a computational redescription of 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—that’s innocuous enough, at least on the face of it. But it is surprising. Why would a two-opsin dichromat *already* have all the resources to take advantage of a third type of opsin, to extract a third dimension of color vision from the activity of a retina with three rather than two cone types? It’s not just that the mechanisms are number-of-cone-types-*general*—you can’t (generally) do the same trick by upgrading from three to four opsins (Jordan et al., 2010). So the result is sometimes resisted.

Before discussing that resistance, let me note something about the dialectic. I’m primarily concerned with computational explanation—how it works, what it assumes, etc.—not with the truth of any particular computational theory. So if it turns out that this finding is mistaken (though I think this is unlikely), it isn’t such a big deal, as long as it is currently accepted. The fact that it is taken to be true by people involved in the giving and justifying of computational explanations is what will shed light on that practice—in a sense, the truth of the finding is irrelevant. But that won’t be totally convincing to most, and it is unsatisfying even if it’s accepted, so let me briefly defend the finding against the two objections I think are most likely to arise. The first objection will argue that computational redescription is not really required in the case above. The second objection will accept that computational redescription is required, but draw from some neuroscientific background to suggest that the redescription actually comes with changes to the redescribed circuits.

*Objection 1.* One might say that the computational structure of the pre- and post-mutation organisms is the same, because their computational structures are not as I’ve described them. One way to make this point is, as briefly considered above, to say that we don’t need to attribute different computations to the dichromat and trichromat. Like I said, this cannot be derived from actual work on dichromacy, but leaving that aside, there is still the problem that the dichromats don’t have three types of cone for the computations to be defined over. Of course, one can gerrymander the system however one likes—all the same circuits are there, so we could say that the circuit that comes to compute L – M in a trichromat descendent was originally performing the same subtraction in the dichromat predecessor, just over different cells. Then there would be no difference in the two organisms’ computational structure. But this is highly implausible. The supposed subtraction in an organism without L cones is just the function M – M, a constant function to 0.[[23]](#footnote-23) That computation is pointless, and more importantly, its stipulation may get in the way of discovering the actual function of the dichromat’s early visual circuits. That is an important research project, undertaken to discover how trichromatic color vision could have been supported by visual circuits that had not developed to support it. Some researchers suggest trichromacy is supported by mechanisms that originally served spatial vision (Shapley, 2009; Wachtler & Wehrhahn, 2016), others suggest new models of dichromacy involving mechanisms that would perform different 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, let alone one like M – M.

The worry was that the dichromat and trichromat might actually have the same computational structure. We’ve just seen that it is not viable to support this by saying we can describe their computational structure in the same way. But it is still possible to say that though redescription is required, it does not correspond to a change in the computational structure of the organisms, because 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;[[24]](#footnote-24) just the opsin expressed in the relevant cone cells has changed. Only a higher-level gloss of the computational structure has changed; the underlying—real—computational structure remains the same.

I gave the bad move a name for a reason. This objection is the one I most frequently hear (from philosophers, at least), and the one I think is most deeply mistaken. The level at which the ‘computational’ structure of the two organisms is the same is *just* *not the level at which the computational explanations I’m discussing are pitched*. The point here is to understand a scientific practice, and that scientific practice does not propose computational explanations at the finer level of grain just floated—it proposes exactly the ones I’ve been discussing, L – M and etc. Here’s another way of putting that point. A level of description at which the pre- and post-mutation organisms are computationally identical would have to be such that either (1) the pre-mutation dichromat *was* doing L – M computations (or something equivalent), or (2) the post-mutation trichromat *wasn’t* doing L ­– M computations. We saw just above that (1) is not an option. But then this objection amounts to (2): a flat denial that the trichromats in question are implementing the computations that are supposed to explain their trichromacy. That is a coherent position, but it needs to be supported in the ways I mentioned in connection with the bad move: it has to be shown that cognitive scientists don’t really attribute L – M computations to the organisms in question, or that they shouldn’t, and neither argument seems to have anything to recommend it, especially not when held up against the ubiquity and success of those attributions.

*Objection 2.* One might accept that the post-mutation organism has a different computational structure than her predecessor, but think there actually *was* a change in her early visual circuits, due to their plasticity rather than genetic instructions (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 precisely *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, Beckmann, Eysel, & Manahan-vaughan, 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., 2009, 2010).

It is possible to make the plasticity suggestion with a caveat: say that the *initial* evolution of trichromacy—in our ancestors and in the mice—was due to plasticity, but in species like squirrel monkeys, where the females are trichromats, there was some selective pressure towards a hardwired setup dedicated to the three-photoreceptor mosaic. Then the squirrel monkey experiment couldn’t be called in to defuse the plasticity worry raised in response to the mouse experiment: the mice developed trichromacy via plasticity, so there were notcommon circuits between the di and trichromats. There is no worry about the time-scale here, because the mice had their entire adolescent periods to develop the right circuits. But, we might say, the monkeys may have developed hardwired trichromatic circuits, so they are not an example of previously dichromatic circuits supporting trichromatic vision—they are just what would happen if the mice had eventually undergone a mutation installing new early visual circuits for trichromacy, instead of letting plasticity handle things.

This is fine as far as it goes, but it reallyhas to reach to get where it’s going. The supposition—that the monkeys are different from the mice in the relevant respects—has little to recommend it, and isn’t taken seriously by any of the scientific literature on the topic that I’m aware of (cf Conway et al., 2010). It is also unclear why there would be any selective pressure for a hardwired trichromatic setup when plasticity served just fine to achieve a similar result. So although this suggestion is a consistent interpretation of the results I described, it seems to have no theoretical motivation except the desire not to allow the computational identity of a circuit to depend on anything external to it, and here we approach the bad move again. If, on the other hand, we let the science speak for itself, the account I’ve given in this section seems to hold up.

There is a related worry: that the circuits themselves didn’t change, but *later*, cortical circuits—ones that the early circuits output to—did change, so the computational redescription of the early circuits is due to internal changes, but not changes to those circuits themselves.[[25]](#footnote-25) This would be interesting (something like a computational version of consumer semantics), and it may be plausible if plasticity later in the cortex happens on a faster time-scale than plasticity in retinal circuits.[[26]](#footnote-26) Perhaps the cortical changes occurred precisely in order to take advantage of a new dimension of stimuli suggested by new kinds of variation in their input from the lower level circuits. Three things should be noted about this suggestion. First is how vague it is. To my knowledge, no experiments suggest that there are mechanisms for the plastic adaptation to different photoreceptor mosaics. To be fair, plasticity is still poorly understood in general, and we shouldn’t expect such a specific theory. But then the suggestion amounts to the claim that there is *some* mechanism at *some* higher level of processing that has *some* kind of relevant plasticity, such that it ends up changing in *some* way that affects the computations performed by the retinal circuits. That’s a lot of ‘some’s, and very little substance, so I see little that can even be responded to in the higher-level plasticity suggestion. Second is that the higher-level plasticity suggestion is not well supported by scientific work, which generally takes for granted that the relevant changes must be to the early circuits. Those who suggest plasticity is responsible for the new color discriminations all (to my knowledge) propose plasticity in the retinal circuits, not in later cortical ones (e.g. Wachtler & Wehrhahn, 2016).

And third, the higher-level plasticity suggestion offers no reason that the move from *three to four* photoreceptors shouldn’t result in tetrachromacy (Jordan et al., 2010). Although the previous two points are revealing in their own ways, I think this one alone is dispositive. If trichromacy results from a plastic reorganization of cortical circuits, the likely mechanism would be a general-purpose one: there could be no selective pressure for mechanisms that reorganize to distil specifically *three* dimensions from the responses of early visual circuits to three photoreceptor types. There could be no selective pressure for this in a dichromat organism because the dichromat couldn’t *use* that mechanism in that way, since it had only two photoreceptor types. What’s responsible for the reorganization would have to be a mechanism that *happens* to reorganize to distil three dimensions from the responses of early visual circuits to three photoreceptor types respond to a third type of photoreceptor, and if this is not some sort of miracle it will be because the mechanism reorganizes *generally* to increase dimensionality in response to variation in its input from early visual circuits. But then that mechanism should increase dimensionality in response to a four-opsin retina as well, and again, we know that it does not (Jordan et al., 2010), at least not generally.

The argument here has not been deductive. These issues are all at the coalface of computational neuroscience—the nature of plasticity especially. But again, even if one is tempted by an alternative explanation (like one of the plasticity explanations) rather than the standard one I’ve given, the fact that the standard interpretation is accepted as legitimate computational theorizing tells us that the practice of computational theorizing allows results like the one I presented. And if that result entails externalism, we’ve established that a system’s computational identity at least *can*—consistently with the project and assumptions of computational explanation—depend on features external to that system. Of course if the standard explanation is correct—and again, I’ve given reasons to think it is—then we’ve established that a system’s computational identity *does* depend on features external to that system.

**5 Externalism**

So far we’ve seen that the computational identity of the neural circuits responsible for color vision depends, in part, on the make-up of the photoreceptor mosaic. The challenge, then, is to find a plausible taxonomic principle that grounds the computational difference between the pre- and post-mutation organisms in the case above. Call them *Pre-M* and *Post-M*. One principle that could deliver this result says that any difference whatsoever in the physical makeup of two systems constitutes a difference in computational structure. This would distinguish Pre-M’s computational structure from Post-M’s, but it would also distinguish Pre-M from a silicon duplicate of her, or a duplicate with just one neuron replaced by a silicon chip—paradigmatic examples of a *non*-computational difference.

One might instead distinguish between Pre-M and Post-M partly on the basis of their differing abilities to exploit 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—how she extracts information from her environment that Pre-M can’t. This is plausible, if sketchy. It is, however, a plausible sketchy *externalist* principle: it appeals to Pre-M and Post-M’s environment and their interactions with it to ground their computational difference. So leave that suggestion aside for now.

The purpose of this section is to pose a challenge to internalists, and sow some pessimism about their ability to meet it. So what can the internalist say about the difference between Pre-M and Post-M? It can’t be merely that there is *some* internal difference between them. But it would be swinging too far in the opposite to invoke a taxonomic principle that says:

T3 Two organisms differ in computational structure if one has the retina of Pre-M and one has the retina of Post-M.

The relevant difference should be stateable in more general terms—it should be what *motivates* the computational redescription, and because of that it can’t merely be an ad hoc redescription of a particular case. A taxonomic principle should make it clear *why* the computational redescription is explanatorily interesting, fruitful, or correct, or at least tell us generally when it is called for. We’re talking about *principles* of taxonomy, not instances of it.

Those two cases suggest some simple guidelines: the taxonomic principle that distinguishes Pre-M and Post-M cannot overgenerate, and it cannot overspecify, because in both cases it will fail to function as we want it to. In the first case it will distinguish organisms that we don’t want distinguished, as in the “any physical difference” attempt. And in the second case it will simply redescribe the case at hand, as in T3, rather than offering a taxonomic principle—a basis on which categorizations can be made.

A taxonomic principle distinguishing Pre-M and Post-M must appeal to their differences, and those could only be (i) a difference in their photoreceptor mosaics, (ii) a difference in some other internal feature—not the actual circuitry, since that’s the same, but perhaps the actual patterns of activity in the circuits—or (iii) a difference in their behavioural profiles. (iii) could be understood as falling under (ii), since the internalist will have to define “behaviour” internalistically, but it is interesting enough to consider separately. I’ll go through each in turn.[[27]](#footnote-27)

 (i) would cause the internalist no end of trouble. As we saw, the fact that there was *some* difference is no help to taxonomy, and the precise difference between Pre-M and Post-M is no help. But nothing in between seems to have much promise either. The specific involvement of a new type of cone—L cones rather than M cones—is no help, because replacing *all* M cones by L cones would have resulted in no computational redescription, 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 (Jacobs, 2008) as a result. The same goes for human trichromats who have 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 M2 opsin with exactly the same light sensitivity as the original M opsin, calling for no computational redescription (that case would be analogous to the case of a silicone substitution). It might also have been an opsin that had the same sensitivity but worked a little slower or faster, likely calling for no computational redescription despite a more consequential change (cf Shevell & Martin, 2017 on some interesting effects of the speed and efficiency of the different types of signals cone cells send).

But perhaps this points the way to a better principle. Isn’t what’s important really the *compound* fact that there is a third type of opsin involved *and* that each opsin has a different wavelength sensitivity profile than the others? 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 organisms have more cone types than dimensions of color vision (M. Neitz & Neitz, 2014), and some human females have four or five types of cone cell (Jordan et al., 2010), but this does not call for a revision in our computational theories of their color vision.

It seems to be that *only* when there is an increase in the dimensionality of vision (e.g. from dichromacy to trichromacy) is redescription required. OK, so add to the compound fact a third conjunct: that the organism becomes a trichromat. The fact that Post-M became a trichromat and the human females above don’t become tetrachromats is important, and seems to be precisely the fact that 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 4) appeals to environmental light sources; in more explicitly externalist terms, this might look something like what I described above: an ability to exploit different features of the environment. It is unclear what the difference between dichromacy and trichromacy would be on internalist terms, if not something like what I discuss under (ii) and (iii).

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 over-specify. (ii) might look more plausible. The internal organization of the organism—its “causal topology” (Chalmers, 2011)—hasn’t changed except in the retina, which was just considered. But beyond the retina the *likely patterns* of activity may have changed. The organism may have certain cells excited more or less often, certain circuits’ activity more 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.

This suggestion threatens to over-generate. Not just any difference in patterns of 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 (e.g. reversing glasses (Harris, 1965)) for significant amounts of time, 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. Recall the discussion in section 3: we don’t want to need new descriptions of the organism’s computational structure for mundane changes to its environment, like an eye patch coming off.[[28]](#footnote-28) So what’s at issue can’t be the actual patterns of internal activity, but something more restricted, like *normal* patterns of activity—a notion that would be difficult to define internalistically, especially in light of the concerns raised the paragraph after next.

Another worry about (ii) comes from the fact that there are enough different alleles of the M and L opsin in human and other old world primate populations that many conspecifics have quite different photoreceptor layers; one expresses S, M1, and L1, and another expresses S, M2, and L2. Aberrant M’s or L’s result in “anomalous trichromacy”—a condition that does not receive an updated computational description (Jacobs, 2008), but does change an organism’s behavioral profile and will, just like the introduction of L cones in the first place, will change post-receptoral patterns of activity. Call the S, M1, L1 organism *Post-M* and the S, M2, L2 organism *Post-M Anomalous*. A taxonomic principle can’t count these two organisms differently, because our computational theories don’t (watch out for the bad move). So, again, the internalist proposal must not just be that when different patterns of activity actually occur in organisms, they have different computational structures.

 What might make some differences matter and others not? Perhaps a similarity metric or set of equivalence classes over internal processes could count Post-M and Post-M Anomalous as the same and Pre-M and Post-M as different, but it’s difficult to see how that metric could be devised non-arbitrarily.[[29]](#footnote-29) It would also likely have to co-classify *Post-M* and *Post-M Faster*—an organism with anomalous M photoreceptors that were faster or more efficient than the old M photoreceptors, but had the same sensitivity profile. This might have far-reaching ramifications on internal activity, and it isn’t clear how the right metric would exclude those ramifications from making a computational difference.

 (iii) is still live: maybe what makes the difference between Pre-M and Post-M is not just their internal patterns of activity, but something to do with their patterns of behaviour, *behaviour* being construed independently of the environment, as something like the movements of the parts of an organism. Patterns of behaviour might be shared between Post-M and Post-M Anomalous, in some sense, but it is not clear what that sense is, at least if we don’t help ourselves to the features of the stimuli they behave towards. Likewise for patterns of behaviour in Pre-M and Pre-M Faster. Maybe there is some similarity metric or set of equivalence classes that can be non-arbitrarily justified, and that count the two as highly similar, but that counts the original Pre-M and Post-M as different. But again, and for the same reasons as above, I’m skeptical that one can be devised. Regardless, the burden is on the internalist to devise one.

I can’t say anything more conclusive here. Internalism is not self-contradictory, and it’s unlikely that it will be outright inconsistent with any observed phenomenon since it doesn’t make direct empirical predictions but merely guides the construction of theories, instructing theorists to pay attention to one set of facts rather than another in their individuating or taxonomizing practices. This is apparent in other discussions of computational externalism as well, e.g. (Horowitz, 2007, pp. 74–75), where internal individuation is dismissed on strikingly brief and inconclusive grounds. A more detailed discussion is better, but the other discussants are right not to spend too much time digging into *every possible* internalist approach, since it is unlikely that any will reveal a contradiction. Rather than a contradiction, 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 can’t be generated from internal features. I’ve just argued that the case I introduced in section 4 presents such a situation. The facts available to the internalist—changes to cone mosaic structure and perhaps post-receptoral patterns of internal activity—offer no clear way to make the taxonomic distinctions we need made.[[30]](#footnote-30) 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 address, and I see no reason to be optimistic about their solving it.

This is a lot like the conclusion reached by a few other computational externalists, but with the important differences that (1) here, it is a real cognitive system that calls for externalist individuation, rather than a toy system (e.g. Shagrir’s (2001) wiring diagram), and (2) I have not had to rely on any notion of representation, or any serious commitments about the notion of computation, to derive externalism. (1) is important because if some explanatory practice calls for externalist individuation, that can’t be proved by showing that a toy system *that does not fall under that explanatory practice* would, if it were to fall under that practice, require externalist individuation. If rocks or bricks or weather balloons fell under the practice of computational explanation, that kind of explanation would look a lot different than it is. It is precisely the fact that those things *don’t* fall under the practice, but cognitive systems do, that makes it interesting.[[31]](#footnote-31) (2) is important because it means that—as I discussed in section 1—the externalism defended here is unambiguously *syntactic externalism*. (2) also leaves us in a position to investigate computational explanation directly, and it allowed us to tie the philosophical question of externalism to methodological questions of real importance to cognitive science. And as I said in section 2 it also makes the argument more general: externalism can’t be rejected by adopting or rejecting any view of representation, or any more committal view of computation.

OK. Challenge posed; pessimism sown. That means we, or at least those of you who are still on board, are externalists—at least insofar as you’re as pessimistic about internalism as I am, and you’re therefore in the market for an externalist account. As much as I’d like to sell you that account here, it would be premature to move from externalism directlyto, say, a view on which computations are individuated by representational contents, as some of the literature does (Horowitz, 2007; Shagrir, 2001). Though I’ve given some hints, a full externalist account will have to be derived from the practice of computational explanation, just like this initial externalist result was, and that calls for another full-length article. For now, I want to conclude by reviewing the motivation and desiderata for an externalist account, and discussing the scientific context in which externalism must function and the kinds of problems it should weigh in on.

**6 Externalist Methodology in Cognitive Science**

I introduced the problem of externalism with the Krakauer et al article on externalist methodology. I want to revisit that issue now, with a few examples, to give a sense of the kinds of interventions an externalist might make into debates in cognitive science.

Krakauer et al’s article raised issues with an internalistic approach—one that takes an organism’s environment to have no influence on that organism’s cognitive structure. As I mentioned earlier, Krakauer and other cognitive scientists generally understand this as a methodological problem rather than a substantive one, but the substantive conclusion reached above clearly bears on the methodological problem. If an organism’s cognitive structure is dependent on its environment, we had better not experimentor theorize like the organism had its cognitive structure intrinsically.

One approach that *does* so theorize is connectomics, at least on some of its more radical presentations. Connectomics is the attempt to understand brain functions by mapping every neuron in the brain, and every one of its synapses with other neurons. We have very 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 (though there is skepticism in some quarters as well). I don’t want to suggest that this appreciation is entirely misplaced, but if externalism is true it must be tempered. Here’s what I mean by that. To discover the computational structure of an organism it will undoubtedly be necessary to understand its internal causal or physical organization. Even theorists who see computational psychology as radically autonomous from neuroscience (Gallistel & King, 2009) agree that questions of implementation—and thus questions about what computations a system is even capable of performing—depend on 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 cognitive structure of the brain (Seung, 2012), or will settle those questions with relatively little input from environment- and task-analysis. That is an internalist approach and, insofar as the cognitive structure of the brain is understood computationally, it is an internalist approach to computation. This is in straightforward contradiction to the conclusions of the previous section. If externalism is true, then the computational structure of the brain cannot be read off from a connectome map, no matter how good that map is. Instead we must look to the environment, and, as I mentioned in connection with the Krakauer et al article above, we need some guide to *what* parts of the environment are relevant, and *how* they are relevant. To be fair, the best work on connectomics does recognize that it cannot replace ecological theorizing (e.g. Morgan & Lichtman, 2013), but if externalism is true, the contribution of the environment must be given close attention, and which features of the environment are relevant will be determined by a fully developed externalist account.

 Another place where externalists can make useful interventions is in the longstanding 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 in conjunction with some reason to care more about one part of the environment than another that this can motivate a position on naturalistic stimuli. Here’s an example. There is some interesting work on stereoscopic depth estimation by naïve subjects as compared to repeat performers of an experiment (Hartle & Wilcox, 2016). In initial experiments, naïve subjects overestimated short distances between objects and underestimated long distances, but subjects who had sat through the experiment multiple times did quite well on both. This is surprising, because the repeat performers are not uniquein training stereopsis—we’re all using it, all the time. So what would make repeat performers so much better than naïve ones? The experiment was originally done with images on a computer screen, designed to provide specific stereoscopic cues. But when Hartle and Wilcox used real objects instead, naïve observers matched the performance of practiced ones. 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 naïve subject couldn’t help making use of.

 So which group of subjects defines the behaviour that we are supposed to be giving a computational theory of? Should we be proposing computations that subserve the naïve subject’s poor performance in the first experiment, or the practiced subject’s good performance? The 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 first-timers were doing things right, or at least naturally, in which case we should presumably take their (shoddy) performance as our explanandum, and 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 naïve subjects’ performance should be set aside, and we should attribute to humans a computational structure that explains the success of subjects who isolated the limited set of cues that underwrite stereopsis.

 Externalism itself doesn’t weigh on one side or the other. You can hold that the computational structure of a system depends on its environment while holding either of the two inconsistent positions that (i) what really matters for computational structure is the processing of cues that isolate the most fine-grained functions possible (as in the practiced subjects), or (ii) what really matters for computational structure is the processing of naturalistic sets of cues (as in the naïve subjects). To be fair, these might not have to be in conflict. It seems natural to say that the processing of naturalistic cues is the ultimate explanandum for cognitive neuroscience, while the processing of artificially isolated subsets of those cues can be understood as one part of the naturalistic process. But there is no guarantee of this: it is possible that the process isolated by the artificially limited cues can’t be understood as a part of the processing of naturalistic sets of cues, e.g. if the processing of naturalistic sets of cues is complex, and not just an ‘adding-up’ of the processing of different types of cue, or if for some other reason the investigations of naturalistic sets of cues and of artificially isolated cues identify processes that carve up the system in incompatible ways. In that case either (i), (ii), or pluralism might be defended.

 These issues arise often in the cognitive science. Going back to color vision, recent technology has made it possible to optically stimulate just one cone cell at a time, and measure the responses of further cells (e.g. Sabesan et al., 2016). This is highly useful for understanding implementation details—as connectomics is—but it is generally taken to also reveal the computational structure of the early visual system (Kling, Field, Brainard, & Chichilnisky, 2019). If externalism is true, that kind of inference from physiological structure to computational structure is far too quick, just as it was for connectomics. 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). So to understand the significance of this research we need to know whether non-naturalistic stimuli can be determinative of computational structure or not. An account of the environmental individuation of computations would go a long way to revealing the true significance of single-cone stimulation studies, and what they are and are not capable of revealing about the visual system.[[32]](#footnote-32)

So, to be clear, externalism itself doesn’t answer the questions I’ve raised here, nor does it settle debates about how the environment should figure into our computational theorizing. But externalism means *that* the environment should figure into our computational theorizing, and a fully worked-out externalism account *would* answer questions about how the environment so figures. It would thereby bring the philosophy of cognitive science into much closer connection with the cognitive sciences and their explanatory practices—a consequence both fields should embrace. In fact, aside from the argument for computational externalism, that has been the main upshot of this article: an approach to computation that takes into account the way computational explanation actually happens in cognitive science—preferably considering this in some detail—can begin to bridge some of the unfortunate gaps between work in both philosophy and the cognitive sciences, and make room for advances on their shared questions.

**Works Cited**

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.

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. https://doi.org/10.1016/j.neuron.2012.12.012

Chomsky, N. (2000). Language from an Internalist Perspective. In *New Horizons in the Study of Language and Mind* (pp. 134–163). Cambridge University Press.

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. https://doi.org/10.1523/JNEUROSCI.4348-10.2010

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. https://doi.org/10.1016/j.visres.2018.06.009

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

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.

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

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, *69*, 103–116. https://doi.org/10.1093/bjps/axw018

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. https://doi.org/10.1101/gr.9.7.629

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

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. https://doi.org/10.1093/cercor/bhy297

Fletcher, S. C. (2018). Computers in Abstraction/Representation Theory. *Minds and Machines*. https://doi.org/10/gd2z3j

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

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

Gertler, B. (2012). Understanding The Internalism-Externalism Debate: What is the Boundary of the Thinker. *Philosophical Perspectives*, *26*, 51–75.

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.

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. https://doi.org/10.1016/j.visres.2016.05.006

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. https://doi.org/10.1073/pnas.1214240110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1214240110

Horowitz, A. (2007). Computation, External Factors, and Cognitive Explanations. *Philosophical Psychology*, *20*(1), 65–80. https://doi.org/10.1080/09515080601085856

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. https://doi.org/10.1113/jphysiol.1970.sp009022

Huberman, A. D., & Niell, C. M. (2011). What can mice tell us about how vision works? *Trends in Neurosciences*, *34*(9), 464–473. https://doi.org/10.1016/j.tins.2011.07.002

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

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. https://doi.org/10.1017/S0952523808080760

Jacobs, G. H. (2009). Evolution of colour vision in mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1531), 2957–2967. https://doi.org/10.1098/rstb.2009.0039

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. https://doi.org/10.1080/0964704X.2014.896662

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

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), 1723–1725. https://doi.org/10.1126/science.1138838

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLoS Computational Biology*, *13*(1), 1–24. https://doi.org/10.1371/journal.pcbi.1005268

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. https://doi.org/10.1109/ISSPA.1999.815735

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. https://doi.org/10.1016/j.conb.2016.09.007

Kling, A., Field, G. D., Brainard, D. H., & Chichilnisky, E. J. (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. https://doi.org/10.1111/ejn.13574

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. https://doi.org/10.1016/j.neuron.2016.12.041

Ladyman, J. (2009). What does it mean to say that a physical system implements a computation? *Theoretical Computer Science*, *410*(4–5), 376–383. https://doi.org/10.1016/j.tcs.2008.09.047

MacLeod, D. I. A. (2010). Into the Neural Maze. In J. Cohen & M. Matthen (Eds.), *Color Ontology and Color Science* (pp. 151–178). MIT Press.

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.

Mancuso, K., Mauck, M. C., Kuchenbecker, J. A., Neitz, M., & 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.

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

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

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. https://doi.org/10.1016/j.jphysparis.2004.03.014

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. https://doi.org/10.1038/nmeth.2480

Nathans, J. (1999). The Evolution and Physiology of Human Review Color Vision: Insights from Molecular Genetic Studies of Visual Pigments. *Neuron*, *24*, 299–312. https://doi.org/10.1016/S0896-6273(00)80845-4

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

Neitz, J., & Neitz, M. (2017). Evolution of the circuitry for conscious color vision in primates. *Eye (Basingstoke)*, *31*(2), 286–300. https://doi.org/10.1038/eye.2016.257

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

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. https://doi.org/10.1111/1468-0017.00109

Piccinini, G. (2008). Computation Without Representation. *Philosophical Studies*, *137*(2), 205–241. https://doi.org/10.1007/s

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. https://doi.org/10.1016/j.conb.2013.10.005

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

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

Rescorla, M. (2013). Against Structuralist Theories of Computational Implementation. *The British Journal for the Philosophy of Science*, *64*, 681–707. https://doi.org/10.1093/bjps/axs017

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice, *8*(12), 1647–1651. https://doi.org/10.1038/nn1606

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). https://doi.org/10.1126/sciadv.1600797

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

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). https://doi.org/10.1007/s11229-018-01921-z

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.

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.

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

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

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.

Wachtler, T., Dohrmann, U., & Hertel, R. (2004). Modeling color percepts of dichromats. *Vision Research*, *44*, 2843–2855. https://doi.org/10.1016/j.visres.2004.06.016

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.

Zele, A. J., & Cao, D. (2015). Vision under mesopic and scotopic illumination. *Frontiers in Psychology*, *5*, 1–15. https://doi.org/10.3389/fpsyg.2014.01594

1. This explicitly blurs the epistemological and metaphysical question: What computations are implemented? How do we know whether a computation is implemented? I think this is fine when the issue is to do with methodology, like it is here... [↑](#footnote-ref-1)
2. See (Milkowski, 2013, Chapter 1) and (Piccinini, 2015, Chapter 1) for two ways of cashing this out. [↑](#footnote-ref-2)
3. Nor does a *narrow* representational account of content establish internalism—as assumed in, e.g., (Butler, 1998)—if externalism does not depend on any assumptions about content. [↑](#footnote-ref-3)
4. The importance of computational structure—rather than computational descriptions of individual events—to computational explanation is discussed in (Ladyman, 2009). [↑](#footnote-ref-4)
5. It is tempting (Dewhurst, 2018; Piccinini, 2015) to say that there is only one *real* syntactic structure—the most detailed description (the three-input description in the example above). But this looks a lot like the the bad move, or at least it will once we come to the objections in section 4. [↑](#footnote-ref-5)
6. The following argument for externalism is sometimes given (Piccinini, 2008; Sprevak, 2010). The implemented algorithms are, as above, mathematical entities, i.e. not part of any physical system. So they are not internal to the physical systems that implement them. But it is a relation to those mathematical entities—the implementation relation—that determines the computational identity of a system. So it is a relation to something external that determines computational identity: externalism is true. But this argument trivializes the question of externalism and makes no advance on the interesting—and important—questions and methodological issues that I discuss momentarily. The external features that have always been at issue are environmental ones, not mathematical ones (but see Gertler, 2012 for discussion of the distinction between physical, communal, and abstract external features in debates over externalism in the philosophy of mind). [↑](#footnote-ref-6)
7. It also seems to be preferred in the debate over externalism about mental states in general (Gertler, 2012). [↑](#footnote-ref-7)
8. Really this is a *schema* for a principle: the “certain relations” need to be specified. [↑](#footnote-ref-8)
9. T1 may run afoul of certain definitions of supervenience, but that shouldn’t pose any problems for the argument to follow, particularly the discussion of (Rust & Movshon, 2005). [↑](#footnote-ref-9)
10. See the papers in (Platt & Ghazanfar, 2010) for good examples of work explicitly premised on affirmative answers to both questions. [↑](#footnote-ref-10)
11. Not to mention that some degree of verificationism is warranted when the concern is with scientific methodology—see the conflation of epistemology and metaphysics noted in the abstract. [↑](#footnote-ref-11)
12. Among many other examples, see (Shepherd & Platt, 2010, p. 526) for an expression of this sentiment in the case of attentional mechanisms in primates, which it is difficult—but essential, Shepherd and Platt argue—to explain in a way that accounts for both results in the laboratory and in natural social settings. [↑](#footnote-ref-12)
13. Though it may be legitimate to ignore some environments or tasks, e.g. non-naturalistic ones. I won’t weigh in on this here, but see section 6 for a brief discussion. [↑](#footnote-ref-13)
14. The limitation in T2 to *certain* environments means that there can be irrelevant environments where computational structure may, for all I have said, change. That might seem to give the S-preferring theorist a response, but S would still be blind to the effects of the environment for all environments within T2’s bounds, and this is precisely the kind of limitation that will prevent S from saying anything interesting about the questions at hand. [↑](#footnote-ref-14)
15. 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, Winawer, Dougherty, & Wandell, 2012). [↑](#footnote-ref-15)
16. For more on opponent operations and their explanatory successes (and shortcomings), see (Shevell & Martin, 2017). [↑](#footnote-ref-16)
17. There is a terminological stipulation here. The L and M opsins are two of many possible forms of the LWS protein (Thoreson & Dacey, 2019). The one that happens to prefer longer wavelengths in a specific organism is generally called the L opsin in that organism, though it would be the M opsin in an organism that also has variant of the LWS opsin that prefers even longer wavelengths. We don’t know which the first trichromat already had and which resulted from the mutation (Wachtler & Wehrhahn, 2016), but switching out the occurrences of ‘M’ for ‘L’ and vice versa has no consequence for the following. [↑](#footnote-ref-17)
18. Could plasticity, rather than genetics, have made a change to the circuitry? Hold that thought for a few pages. [↑](#footnote-ref-18)
19. See, among many other articles, (Conway et al., 2010; Huberman & Niell, 2011; Jacobs, 2008, 2009; Jacobs & Nathans, 2009; Kóbor, Petykó, Telkes, Martin, & Buzás, 2017; Mancuso et al., 2009, 2010; Mollon, 1984; J. Neitz & Neitz, 2011; M. Neitz & Neitz, 2014; Shapley, 2009; Wachtler et al., 2004; Wachtler & Wehrhahn, 2016). [↑](#footnote-ref-19)
20. Except where dichromacy is studied as a deficiency in an otherwise trichromat species. In that case it is natural to assume the dichromat has the trichromat’s computational structure, and see what went wrong with it to explain instances of dichromacy. But work on dichromacy itself, e.g. in mice, neither makes or supports the assumption that if only the retinal mosaic changed, we could posit no change to the computational structure of later circuits. [↑](#footnote-ref-20)
21. 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 any different from the one where *across the board* dichromacy is turned into across the board trichromacy, which is the situation I’ll discuss. [↑](#footnote-ref-21)
22. Squirrel monkeys have this in common with the rest of the new world monkeys, one important exception being the 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). [↑](#footnote-ref-22)
23. If it were a center-surround opponent cell it might be useful for spatial vision, but then it wouldn’t be a constant function to 0, it would be a center–surround opponent computation—*not* the simple computation M – M (see Wachtler & Wehrhahn, 2016). [↑](#footnote-ref-23)
24. At least we can stipulate that they don’t. The connections are generated stochastically so they will differ between any two organisms, but we can set this aside to get to the heart of the objection. [↑](#footnote-ref-24)
25. A related suggestion is that the early circuits don’t call for redescription, but the later ones that change *do*. I’ll set this aside because it is another instance of the bad move, standing in conflict with the actual theories we give of trichromat and dichromat organisms, all of which involve the previously discussed opponent computations being performed in the retina and LGN. [↑](#footnote-ref-25)
26. In principle this is hardly different from the case I’m arguing for, where an internal change to the retina affects the computational identity of the circuits that come after it. In both cases an internal change calls for a redescription of another part of the system. But the argument in section 5 goes much more smoothly if we can restrict our attention to the ways that the cone mosaic can affect computational identity, and don’t also have to consider the complicated variety of ways that later processing might affect it. [↑](#footnote-ref-26)
27. This is why it was important to rule out plasticity in the previous section. It might be the case that there is no internalist principle that could distinguish Pre-M and Post-M successfully even on the basis of changed circuits (either the computation-performing ones or later ones), but this would call for an intricate and very subtle argument. Dealing with just (i)-(iii) will be more straightforward. [↑](#footnote-ref-27)
28. Two notes. First, the actual computational operations performed by the system may change in response to deprivation or reversal. The question is whether the computational *structure* of the system changes. Your calculator and mine do not differ in computational structure simply because yours is adding 2 to 7 and mine is adding 5 to 8. And your visual system and mine do not differ in computational structure just because you’re looking at something different than I am. They differ in computational structure when they call for different computational theories (see section 2), and it is implausible that donning an eye patch has that effect (again, see the discussion of the “any arbitrary stimulus” condition in section 3).

Second, of course under some circumstances an eye patch, or something more severe like optical ablation, may be enough to change computational structure—e.g. this is plausibly the case when deprivation begins at birth and prevents development (Hubel & Wiesel, 1970). But this is a situation where the circuits involved actually change during development—not at all the case at hand. [↑](#footnote-ref-28)
29. Non-arbitrariness matters because there will always be a grouping that groups the right things together, since every grouping exists. The point is to find a rationale for (i.e. a taxonomic principle delivering) the right one. The mere existence of the grouping sheds no light on computational explanation. [↑](#footnote-ref-29)
30. As I mentioned in a previous footnote, isn’t even clear that changes to latercircuits, which were discussed in section 4, *or even changes to the relevant circuits themselves*, would suffice for a taxonomic principle (if they had actually occurred). This is one result of the choice of definition in section 3. If an internal change is necessary to license a new computational description, but the justification of the redescription must refer to the environment, we have externalism, not internalism. [↑](#footnote-ref-30)
31. This is to say nothing at all about the question of *why* they don’t fall under that practice, or how we should characterize it so that they don’t—both substantive questions in the philosophy of computation. [↑](#footnote-ref-31)
32. (Shepherd & Platt, 2010), which I mentioned in section 3 in connection with the Krakauer et al article, is another good example of work for which both issues are relevant. [↑](#footnote-ref-32)