How to be a realist about computational neuroscience

Danielle J. Williams
Washington University in St. Louis
danielle.williams@wustl.edu

Abstract Recently, a version of realism has been offered to address the simplification strategies used in computational neuroscience (Chirimuuta, 2023; 2024). According to this view, computational models provide us with knowledge about the brain, but they should not be taken literally in *any* sense, even rejecting the idea that the brain performs computations at all(computationalism). I acknowledge the need for considerations regarding simplification strategies in neuroscience and how they contribute to our interpretations of computational models; however, I argue that whether we should accept or reject computationalism about the brain is a separate issue that can be addressed independently by a philosophical theory of physical computation. This takes seriously the idea that the brain performs computations while also taking an analogical stance toward computational models in neuroscience. I call this version of realism "Analogical Computational Realism." Analogical Computational Realism is a realist view in virtue of being committed to computationalism while taking certain computational models to pick out real patterns (Dennett, 1991; Potochnik, 2017) that provide a how-possibly explanation without also thinking that the model is literally implemented in the brain.

0. INTRODUCTION

Computational neuroscience takes the perspective that the brain solves a variety of computational problems. Computational models are meant to "provide intuition" about how the brain, "a complex, nonlinear dynamical system with feedback loops," solves those problems (Sejnowski, 2015, pg. 481). While there is a wide variety of computational model types, what they all have in common is the reliance on necessary modeling strategies such as abstraction and idealization. Abstraction is most generally understood as the omission of details (Cartwright, 1999; Frigg, 2006; Nowak, 1992; Strevens, 2008; Weisberg, 2013) or as providing a "detail poor" representation (Levy, 2021). Indeed, the brain is incredibly complex, dynamic, and evolving—a system far too complicated for every feature to be included in the model. Thus, computational models are simplifications in the sense that they leave out key features of the target system. Idealization, on the other hand, involves the intentional misrepresentation of features of the target system

(Levy, 2021). The introduction of idealized features incorporates known falsehoods into the model that distinguish it, in important ways, from the way the world actually is, a further simplification strategy.

The use of idealization throughout the sciences has contributed to a long-standing debate about what kind of knowledge scientific models (or theories) can provide, whether we should be committed to the existence of the object posited with them, and whether we should interpret the resulting, idealized models as literal representations of the way the world is (Psillos, 1999). To use van Fraassen's (1980) words, an initial naïve framing of the position of scientific realism is that "the picture which science gives us of the world is a true one, faithful in its details, and the entities postulated in science really exist: the advances of science are discoveries, not inventions" (van Fraassen, 1980, pgs. 6-7). For the realist, truth plays an important role in the position, along with formulating what it means to accept a theory. As van Fraassen points out, though, the naïve statement is *too* naïve, and instead, the realist is better understood as holding the position that science *aims* to give us, in its theories, a literally true story of what the world is like and accepting a scientific theory involves the belief that the theory is true (van Fraassen, 1980, p. 8). This view takes realism to accept that science aims at truth (even if their theories are not literally true) while also believing, to varying degrees, that those theories are true.

Chirimuuta (2016, 2023, 2024a) has recently formulated a version of realism that she calls "haptic realism," which argues that computational models, in particular, should be interpreted as having only a *constructed* similarity between brains and computations where computation provides scientists with a useful simplification of the brain. The "haptic" naming of her view reflects Chirimuuta's desire for a shift in metaphor from one that associates "knowing with vision" to one that takes knowledge acquisition to be something more akin to *touch*—where scientists learn about the world by interacting with it rather than passively observing it. A resulting implication of this view is that it rejects the idea that the brain performs computations—a rejection of the computationalism thesis.

In this paper, I take lessons about the simplification strategies in neuroscience seriously, especially as Chirimuuta formulates them with respect to computational neuroscience. Alternatively, though, I argue that we can respect the simplification strategies without also rejecting computationalism, meaning we can be a realist in the sense that the brain literally does perform computations while also accepting that computational *models* do not say literally true things about the computations performed by the brain. I call this version of realism "Analogical Computational Realism" (ACR). ACR is a realist position in that it is committed to computationalism: it takes as true the idea that the brain performs computations. But it prescribes that computational *models* should not be interpreted literally in the sense that we should not take them to provide a literally true story about which computations the brain performs or how the brain performs

those computations.[1] Instead, we should take specific computational models as picking out "real patterns" (Dennett, 1991; Potochnik, 2017) that can serve as "how-possibly" explanations regarding computation in the brain. The "analogical" part of the view is understood in terms of a comparison between two ways of investigating the brain. One end of that analogy involves providing a computational model that serves as the computational theory, while the other end of the analogy involves investigating the underlying material system that is responsible for the phenomenon. The two ends of the analogy are tethered to each other through the mind-independent fact that the brain performs computations.

This paper proceeds as follows. Section 1 provides a brief overview of scientific realism as compared to the formulation of haptic realism. I think that haptic realism provides a good reason for thinking more deeply about what kind of knowledge computational models in neuroscience provide; however, I argue that we do not need to reject (as haptic realism does) the computationalism thesis. This situates the truth of the theory with the computationalism thesis while taking the computational model as providing an analogy for how and what the brain computes. Given what we know about the nature of computational modeling strategies, this position provides the strongest realist path forward when it comes to understanding what computational models can tell us about computational brain processes. In Section 2, I motivate ACR in the same way that Chirimuuta motivates haptic realism by describing abstraction and idealization in computational models. However, I reach a different conclusion about the status of those models. In Section 3, I provide a detailed account of ACR. In this section, I argue that a specific class of computational models can be understood as tracking real patterns (Dennett, 1991; Potochnik, 2017) and that those real patterns provide a how-possibly explanation of computational processes in the brain. In Section 4, I discuss how ACR allows for a model pluralism—a good-making feature that can, however, be managed or tamed. I close with two ways in which we might begin to tame ACR: model comparison strategies and criteria-based model selection.

1 SCIENTIFIC REALISM

A long-standing debate in the philosophy of science centers around questions regarding what kind of knowledge scientific models (or theories) can provide, whether we should be committed to the existence of the object posited within the theory and whether we should take the models as a literal interpretation of the

[1] This view is neutral about what type of computational system the brain is, i.e., whether it performs digital, analog, some *sui generis* type of computation, or a mixture. ACR is compatible with the view that only some of the brain performs computations and that computational processes may extend beyond the brain. For simplicity, I will talk about how computational models relate to the brain, but this is not to exclusively confine computational processes to the brain. For example, computational processes may involve features that are not properly considered a part of the brain, such as the cochlea or aspects of the somatic nervous system.

way the world is (Psillos, 1999).[2] Modern science has transformed the way we investigate, understand, and model the world. Its tools and methodologies have allowed us to model and understand features of the world that are otherwise invisible to our perceptual systems—entities such as DNA molecules, electrons, and electromagnetic waves. Because scientific theories posit unobservable entities that are often underdetermined by the data (among other concerns), we can ask whether we should be committed to what scientific theories tell us about the world.

The most powerful reason for being a realist about scientific theories is the 'no miracles argument' which says that realism "is the only philosophy that doesn't make the success of science a miracle" (Putnam, 1975, pg. 73). This is the idea that taking our best scientific theories as true is the only hypothesis that does not make the astonishing predictive and explanatory success of science a mystery. Two additional reasons for adopting the realist position include the idea that unobservables described by our best scientific theories are indispensable in the success of the theory itself. If aspects of the theory are essential to the novel predictions that are made by the theory, then they are worthy of a realist commitment (Kitcher, 1993; Psillos, 1999), and finally, we should take a realist position because of our ability to causally manipulate unobservable entities (Hacking, 1982 and 1983; Cartwright, 1983; Massimi, 2004).

Scientific realism can be and has been formulated in many different and nuanced ways. To a first approximation, we can understand realism as having two general positions, each with its array of nuanced sub-positions. The first general position characterizes science in terms of epistemic achievements constituted by scientific theories and models. On this approach, realism is a position that concerns the epistemic status of theories or some component thereof. What these approaches have in common is a commitment to the idea that our best theories have the epistemic status of yielding knowledge of aspects of the world, including those aspects that are unobservable. The second general position understands realism in terms of the epistemic aims of scientific inquiry. These positions are characterized in terms of what science aims to do—that it aims to produce true descriptions of things in the world (or approximately true descriptions). Following van Fraassen (1980), I will take this second line, focusing on the aims of science when it comes to understanding the role that computational models play in theorizing about the computational processes in the brain (Section 3).

1.1 Haptic realism and computational neuroscience

Some have proposed that we think of scientific practice as being influenced by the practices and perspectives of human agents, which makes observation and theorizing a perspectival project rather than a

[2] In this section I use the terms 'model' and 'theory' interchangeably in some places. This is merely to facilitate the explication of scientific realism; it is not to commit scientific realism (or myself) to the idea that models and theories are the same kinds of things.

practice that captures an objective reality. This view of scientific practice is called perspectivalism. Like standard realism, perspectivalism can be specified in different ways. For example, Giere (2006) proposes a version that maintains the idea that, like the human visual system, instruments are sensitive only to a particular type of input, and they are not perfectly transparent—meaning no amount of calibration will ever eliminate the contribution that the instrument contributes to the outputs of the measurement. Because of this, Giere argues that "all theoretical claims remain perspectival in that they apply only to aspects of the world and then, in part *because* they apply only to some aspects of the world, never with complete precision" (Giere, 2006, p. 15). Another version of perspectivalism is proposed by Massimi (2023), who maintains that what makes a scientific practice a scientific perspective is that the knowledge it produces is situated: the knowledge is produced by a community at a time within a cultural context within the boundaries of the resources available to them rather than through some independent reality (Massimi, 2023, pg. 6).

A new version of perspectivalism is Chirimuuta's recently proposed "haptic realism" (Chirimuuta (2016, 2023, 2024a). The "haptic" naming of the view comes from her rejection of what she calls "the visual metaphor" of scientific investigation. She argues that the relationship between the tools used in neuroscience and the world under investigation conceptualizes the relationship between the scientist and the world in terms of a visual metaphor where the tool pictures the target.[3] She takes this metaphor as associating vision with knowing, which frames the world under investigation as something that can be presented directly to us upon observation. This gives the impression that we can apprehend some objective structure of the world through the use of our tools. She says that to truly understand perspectivalism, then, requires abandoning the vision metaphor and instead adopting one that respects how scientists learn about the world through "tinkering and interacting with it" (Chirimuuta, 2016, pg. 755). Haptic realism is proposed as a general version of perspectivalism but has been specifically applied to the context of neuroscience and computational modeling practices (Chirimuuta, 2024a). This application of the view centers on the idea that knowledge acquisition in neuroscience is an interactive process that involves the use of highly idealized computational models that are used as tools to understand the brain (Chirimuuta, 2023, 2024a).

Chirimuuta (2023, 2024a) has argued that traditional scientific realism has been developed with an "eye toward physics" and has then been retrofitted for the special sciences (neuroscience in particular). This use "[centers] around questions of the existence of unobservables, one that occurs naturally for physics and chemistry, but seems irrelevant to many branches of biology" (Chirimuuta, 2023, pg. 1). One of her

[3] She cites Teller (2001, pg. 393) who makes this same point.

motivations, then, is to develop a realism for a scientific context where unobservables are not generally posited.[4] When thinking about how traditional realism is "physics-centric," Chirimuuta says the following:

> …what's peculiar about physics is that it investigates the most simple stuff around: nonliving matter, not undergoing chemical reactions. The stuff under consideration for the physicist is quite *homogenous* (particles of a certain type are all the same, unlike members of a biological species), *unchanging* (not subject to growth, plasticity, and age), and *insensitive* to its surroundings. Neurons, the 'elementary particles' of neuroscience, are very much unlike this (Chirimuuta, 2023, pg. 3).

The idea is that the objects of neuroscience are much more complicated than the objects of physics because, unlike physics, the brain lacks fixed targets. What makes the brain different is that it is constantly changing itself in response to how things are in the body and the rest of the world such that any representation in neuroscience "will be no more than a rough, rigid caricature of a protean object that can never be fully characterized" (Chirimuuta, 2023, pg. 5). To this point, Chirimuuta discusses the role of idealization in neuroscience and argues that "traditional realism is guilty of neglecting the difference between the "idealized world" within the theory and the more complex reality that the theory represents. Idealization creates a gap between the theory and the actual brain where the "gap" is understood as the space between how the world is and what the model says about it. Relating this to neuroscience, she says, "the objects of neuroscience are much more complicated than the objects of physics, which means the gap between the object of investigation and the idealized representation will be even larger" (Chirimuuta, 2023, pg. 4).[5]

Using neuroscience as her science of interest, Chirimuuta argues against the metaphysical stance adopted by the traditional realists. The metaphysical stance is the commitment to the mind-independent existence of the world investigated by the sciences. Mind-independence, to Chirimuuta, essentially means

[4] One thing to consider here is that scientific realism is committed to the claim that the entities posited by our best theories exist, whether they are observable or unobservable. But if your science doesn't posit unobservables, then it is not clear what would motivate you to be a scientific realist rather than an empiricist (who says that the entities we can observe unaided exist). So scientific realism isn't just a claim about unobservables, but it does include a commitment to unobservables, which brings extra ontological baggage. It's not clear what the motivation is to be a scientific realist about a particular science if it doesn't posit unobservables, but the lack of unobservables doesn't prove scientific realism wrong. This helpful comment was provided by Zoe Drayson on an early draft of this paper.

[5] Cartwright (1983) critiques the use of idealization in physics in much the same way that Chirimuuta does, arguing that the use of idealization in physics puts pressure on the traditional realist. Cartwright points out that for "the idealization [to be] of use, when the time comes to apply it to a real system, we had better know how to add back the contributions of the factors that have been left out" (Cartwright 1983, pg. 74). While Cartwright seems to be describing something like abstraction in this quote, she references the use of non-real properties later indicating that her target is idealization as described here (Cartwright 1983, pg. 102). Nonetheless, her point still stands: if the details either do not matter or we know how to treat them, then there is no challenge to traditional realism. (Cartwright 1983, pg. 74). However, if we do not know how to treat them, then the traditional realist is in trouble. Chirimuuta can be understood as making a similar point about the computational models in neuroscience. As a reviewer helpfully pointed out, this means the issue Chirimuuta (and I) target in neuroscience arguably exists in physics as well.

human independent. Human independence, she says, takes scientific knowledge to somehow transcend the relatedness of knowledge to the scientist's *circumstances of discovery* and *technological motivations*. Given the trade-off between precision and accuracy in modeling practices, vindicating the metaphysical commitment becomes difficult "once we appreciate that even the best representations of what there is in the brain are imprecise and simplified…" (Chirimuuta, 2023, pg. 9). This doesn't mean, however, that the representations or models in neuroscience cannot yield understanding of their target in some way.

Chirimuuta's haptic realism is placed into the broader landscape of the philosophical and neuroscience literature in her book, *The Brain Abstracted* (Chirimuuta, 2024a). Here, she carefully spells out the history of simplification in the brain sciences in a robust and enlightening way. In this project, she makes the following argument:

> "…The mathematical structures that make the brain intelligible to scientists, as an organ whose function is to process information, are to some extent imposed onto the neural system by the scientist and should not be taken as straightforward discoveries of mathematical forms inherent in the system. Since, by hypothesis, neurocomputational models are not discoveries on the inherent computational capacities of the brain but are abstract and idealized as any other models in science, an analogical interpretation of these models is more appropriate than a literal one" (Chirimuuta, 2024, pg. 106).

This analogical interpretation stands in contrast to a literal interpretation of neural-computational models. Instead, she argues that we should understand the model as identifying a pattern that is based on the judgment of the scientist and not a forced choice that comes out of the data (Chirimuuta, 2024, pg. 111). Chirimuuta calls these patterns "ideal patterns" and argues that they do not exist independently of the scientist's activities, data processes, or theoretical choices. Given this, her interpretation of the knowledge provided by computational models is that they do not convey essential features of the neural-cognitive system, but rather, they provide an idealized pattern of those processes. Ideal patterns are the regularized products of a series of simplifying procedures such as abstraction and idealization (Chirimuuta, 2024, pg. 129). They are "ideal" to mark them out from the "real patterns"—or actual patterns that exist "out there" independently of the human investigator.

An ideal pattern is contrasted with Potochnik's (2017) use of "real patterns." Potochnik proposes that real patterns are the target of scientific models, an adapted version of Dennett (1991).[6] Consider Dennett's "bar code" example that includes a pattern of 10 rows. Each image contains 90 dots, and every row within the image includes a variation of 10 black dots followed by 10 white dots:

[6] It is an adapted version because Potochnik considers patterns in phenomena while Dennett considers patterns in data.

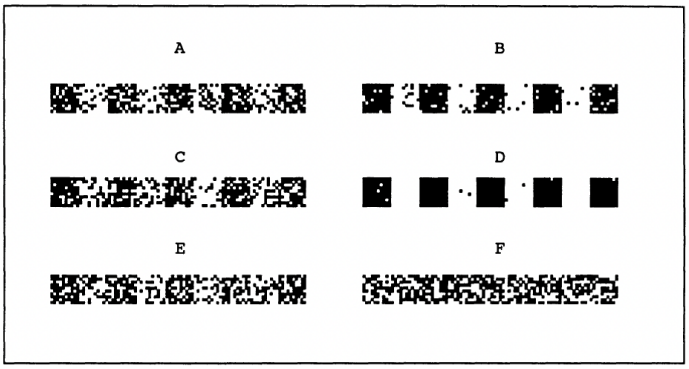

Figure 1. Dennett (1991) uses these objects to articulate the idea of a "real pattern." Used with permission from the *Journal of Philosophy*.

When describing the image, Dennett says that in some respects, A-F all display different patterns. But, in another sense, A-F display the same pattern because "they were all made by the same basic process, a printing of ten rows of ninety dots, ten black dots followed by ten white dots, etc." (Dennett, 1991, pg. 31). Because the overall effect is to create 5 equally spaced black squares, Dennett argues that the pattern shared across the different representations "is a real pattern if anything" (Dennett, 1991, pg. 31). Potochnik takes this idea and applies it to the explanation of a phenomenon and argues that we can depict a system by citing a pattern in much the same way. In doing so, we can indicate the level of deviation the phenomenon has from that precise pattern, or we can refrain from mentioning the amount of deviation, as all that matters is whether a system "embodies the pattern" in question (Potochnik, 2017, pg. 27).

For Potochnik, the patterns embodied in the system are really there in that they actually exist within the system. Whether a pattern is revealed depends on what we can and choose to observe about a system, along with what we can and choose to represent about that system (Potochnik, 2017, pg. 28). For Chirimuuta, however, these patterns are *not* really there. Instead, the patterns are ideal in the sense that the patterns come as a product of the scientists tinkering with the system along with the tools they use; they are not "out there" to discover. The difference between these views is that Potochnik argues that the patterns come from the system itself, while Chirimuuta argues that the patterns come from the scientists and do not exist, in any literal way, within the system under investigation.

One important aspect of Chirimuuta's view that is made explicit in her 2024 book is the idea that the brain does not literally perform computations (a denial of computationalism). Instead, she argues that the relationship between the brain and computers should also be understood as an analogy in that computational modeling does not "warrant the conclusion that the neural systems themselves compute the functions specified in the models, or that the brain itself is literally a computer" (Chirimuuta, 2024a, pgs. 91-92). According to Chirimuuta, "the analogical interpretation is a doctrine of restraint: it declines to infer from the success of the computational approach in neuroscience that the brain really is a computer—an

organic device performing calculations to which the neurocomputational models provide a closer or wider approximation" (Chirimuuta, 2024, pg. 112). This way of thinking about computational models denies the inference from the success of the computational approach to the metaphysical thesis that the brain literally performs computations.

Chirimuuta takes the denial of computationalism to be a positive attribute of her view in that it eliminates the need to take on "difficult metaphysical commitments and philosophical challenges" (Chirimuuta, 2024, pg. 112). She says that taking computational models literally subjects the realist to triviality objections because the realist will be required to appeal to a theory of computational implementation. Here, Chirimuuta is referring to a philosophical theory of computational implementation that specifies the relation between a formal computational structure and a physical system (Putnam, 1988; Chalmers, 1995; Williams, 2023). The problem with implementation theories that rely on mapping the formal structure to the physical system (as most of them do) is that, in their simplest form, they are subject to a triviality objection: we can map countless computational structures onto countless physical systems, capturing physical systems that obviously do not perform computations. The challenge, as Chirimuuta describes it, is to give a theory of implementation that avoids triviality while also showing how the computational level of explanation is somewhat autonomous from the implementation one. This means that there is a problem for the realist who accepts computationalism on two fronts: specifying implementation without being subject to triviality and saying why computational explanations can be given independently of details about the physical system (Chirimuuta, 2024, pg. 113).

I take Chirimuuta's development of haptic realism to be doing something quite important when it comes to considering whether or how we should be a realist about computational models in neuroscience. I take this lesson seriously as a caution for thinking that computational models provide us with literally true descriptions of what the brain is doing. In light of the heavy use of abstraction and especially idealization, thinking about computational models strictly in terms of whether they can tell us truths about the brain is the wrong way to approach questions about how these models *relate* to the brain. However, while I acknowledge the contribution that haptic realism makes to thinking about realism in the context of computational neuroscience, I argue that we can be mindful of this feature of the models without denying that the brain literally performs computations. In other words, we can deny that computational models say literally true things about the brain without also denying computationalism. An upshot of reinstating computationalism is that we can maintain a version of realism that takes the computational aspect of the computational models as tracking actual features of the brain in some respects—we can support a stronger realist position.

In the following sections, I argue for a version of realism that takes the role of simplification (especially idealization) when it comes to computational modeling seriously without taking the further step

of denying computationalism. Instead, I argue that whether the brain counts as a computational system should be left up to a metaphysical theory of physical computation.[7] Note that this is in opposition to what Chirimuuta says if we take theories that address the nature of physical computation to be exhausted by mapping views that target the implementation relation:

> … No theory of implementation is uncontroversial, and appealing to such a theory cannot by itself make the case for a formal realism over formal idealism. One positive argument for formal realism would be to say that if the computational description is a useful simplification—a good analogy—it must be that it does a good job of capturing the structure of the target system. That, then, is reason to think that the system is literally computational. Conversely, if the target system is not literally computational, then the computational approach must provide a poor simplification and a misleading analogy. But this argument simply assumes that models work—provide useful simplifications—to the extent that they faithfully represent structures that exist in the target system, an assumption at odds with so much work in the philosophy of science on modeling, abstraction, and idealization (Chirimuuta, 2024, pg. 114).

Mapping views, however, are not the only way to understand computation in physical systems. Moreover, theories that address the nature of physical computation may garner some support in light of the success of computational modeling, but the theory itself is not established based on this observation. Thus, I argue that we can still look to the work on physical computation to underpin the computationalism thesis independently of the success of computational models. This strategy pulls issues about whether the brain counts as a computing system apart from questions about whether the brain implements a given computational model. That is to say, not all theories addressing the nature of physical computation propose that determining which physical system performs computations requires that we map a computational description to a physical system. Put within the context of neuroscience, not all theories that address questions about whether the brain performs computations require that we map a computational model onto the brain. This is consistent with what Chirimuuta calls a "formal realism"—or a traditional realism—However, this version is not at odds with work in the philosophy of science on modeling, abstraction, and idealization, as we will see, because the view does not demand that we take computational models as providing a literal interpretation of the computational structure of the brain nor does it require that we ascribe the computations performed by the model to the brain. Before developing this view in detail, I will motivate my view, like Chirimuuta, by discussing the role of abstraction and idealization in computational modeling practices.

[7] To say this is just to say that an answer to the individuation question, which systems count as computers and which systems do not, is a question that can be addressed independently of how we should formulate scientific realism. This is not to endorse any particular theory of physical computation.

2 ABSTRACTION AND IDEALIZATION IN COMPUTATIONAL MODELING

Computational neuroscientists use mathematical models, computer simulations, and statistical analysis to understand the workings of the brain, nervous system, and behavior with the goal of explaining, in computational terms, how brains generate behavior. The resulting cognitive computational models are used to simulate cognitive functions such as perceptual tasks, attention, decision-making, and more. In these cases, computational models provide a highly simplified and idealized way of understanding computational information processing during a task (for an example see Pouget & Sejnowski, 2001).[8] While there are many different types of computational models used across neuroscience, what they all have in common is that they rely on necessary modeling strategies such as abstraction and idealization.[9] [10]

To a first approximation, abstraction is typically understood as a lack of detail (or a stripping away of features) in a model (Cartwright, 1999; Frigg, 2006; Nowak, 1992; Strevens, 2008; Weisberg, 2013), while idealization is understood as the introduction of distortions into a model (Jones, 2005; Godfrey-Smith, 2009a; McMullin, 1985; Laymon, 1995; Weisberg, 2007). Levy (2021), though, argues for a conception of abstraction that is not tied to truth.[11] "Abstractness," under this view, involves providing a "detail-poor representation" (Levy, 2021, pg. 5858). Where this view differs from others is that whether a representation counts as more or less "abstract" depends on its relation to other representations rather than the true representation that includes all features of the system (Levy, 2021, pg. 5859).[12]

The omission of details, which gives rise to a detail poor representation is not merely a modeling choice. Rather, it is a necessary part of the computational modeling process. For example, because the brain takes in inputs across several dimensions and involves many complex and changing processes, a model that includes every feature in the target system will fail to run. The idea that we cannot include all variables in a model is known as "the scaling problem" (Churchland & Sejnowski, 1992, pg. 125). If a computational model is scaled up to incorporate all of the relevant dimensions present in the target phenomenon, then in

[8] This is not to say that information processing and computational processes are the same thing. If we are to take the computationalism thesis seriously, then mere information processing is insufficient for ascribing full-blown computational processes to the brain.

[9] In this section I use 'representation' with no specific theory of representation in mind. While I specifically target computational models, representations may also include texts, graphs, and other types of physical models, all of which may make use of abstraction and idealization.

[10] Varying degrees of abstraction and idealization are used depending on what is being modeled along with the explanatory aims of the scientist deploying the model.

[11] As Levy points out, 'abstraction' sometimes refers to a feature of representations, namely, a poverty of detail. On the other hand, 'abstraction' is also sometimes described as the "leaving out" or "omission" of detail. The latter way of thinking about abstraction involves understanding it as a *process* (literal omission of details such as values), while the former takes abstraction to be a product. It is this former way that Levy investigates in his paper (Levy, 2021, pg. 5858). Because I am interested in the model itself, as an abstraction, I adopt Levy's formulation.

[12] In this context, "abstract" is understood as "abstract-as-omission" rather than "abstract-as-abstracta." For a discussion on the difference, see Williams (*forthcoming*).

some cases, the model will perform too poorly to be usable, or it will fail altogether. Importantly, though, more goes into generating the model than the omission of details. The model is not just "what's left" after details are omitted.

Idealization during modeling is a deliberate misrepresentation of what the world is like.[13] Idealization, as opposed to abstraction, is related to truth since what it says is known not to be true of its intended target (Levy, 2021, pg. 5861). The way that truth operates when it comes to idealization is that we take what is already known about a system and we intentionally misrepresent it. For a representation to be deliberately false in this way, we are required to know the actual truth about the target system. While abstraction has to do with the choices about including and excluding details (either by choice or because we do not yet know how to fill in those details), idealization is a matter of misrepresentation in that it concerns the relation between the representation and the way that the world actually is (Levy, 2021, pg. 5862).[14] Just like abstraction, idealization is also a necessary strategy for modeling brain processes in neuroscience. This is what allows for the mathematical specification of complex and dynamic phenomena. This is just a feature of modeling the world in that "all theories, even the best, make idealizations or other false assumptions that fail as correct descriptions of the world" (Wimsatt, 1987, pg. 1).

One reason for the introduction of idealized components into a computational model sometimes has to do with "the segmentation problem" (Churchland & Sejnowski, 1992, pg. 83). This is the problem of how sensory stimuli are presented in nature versus how they can be inputted into a model. Real-world inputs do not come into nervous systems informationally packaged into separate batches labeled for separate problems. Instead, real-world inputs are noisy and oftentimes ambiguous. Consider visual information. The nervous system is tasked with the immediate job of separating information concerning motion, shape, size, distance, etc., coming into the retinas. To accomplish this task, the nervous system must separate objects in the environment by distinguishing relevant edges and boundaries; it must determine which objects are relevant to the moments and at what distance the objects are in relation to both the retinas and other objects. But often, determining where these boundaries begin and end is not a straightforward task. This means that the inputs into the computational model must be simplified through idealization as well.

---

[13] It is possible to construe abstraction as a kind of misrepresentation because it represents the target as lacking some of its details. However, when drawing out the distinction between abstraction and idealization, the difference is in how the *features* of the target system are represented. So, in the case of abstraction, *features* are left out whereas in idealization *features* are misrepresented.

[14] Weisberg (2013) distinguishes between three types of idealization: Galilean, minimalist, and multiple models. Galilean idealization is the practice of introducing distortions into a model with the goal of simplifying. Minimalist idealizations involve constructing and studying models that include only the core causal factors while multiple models idealization involves building multiple related, but incompatible models. The sense in which this paper understands idealization is in Weisberg's Galilean sense (Weisberg, 2007, pg. 99).

Abstractions and idealizations are simplifications. However, this does not mean that the resulting models are inherently simple. They are simplifications only in the sense that they are strategies to make the target system manageable such that we can use formal tools to generate a usable model of the complex, dynamic system. Many philosophers have pointed to the fact that the world is not simple, that it is disordered, complex, and dappled (Dupre, 1993; Bechtel & Richardson, 1993; Cartwright, 1999; Wimsatt, 2007; Mitchell, 2012). The brain is perhaps one of the most complex systems scientists investigate. It includes around 85 billion neurons along with a similar number of glial cells. Additionally, each neural cell has its own complexity, relying on complicated chemical and electrical signaling processes, with synaptic connections that range from a few hundred to thousands. These cells have connectivity differences, functional differences, participate in different processes, and differently in the same process, they come in different sizes and shapes, creating a thicket of connections across the brain. To model a system of such complexity, it is necessary to leave some features out and also simplify other features by misrepresenting them. So, while the models rely on necessary simplifications, this is not to say that the model itself is simple; it is just simpler than that target system.

2.1 Ordinary computational models, computational explanations, and physical computation

Before moving on, some preliminary clarifications are needed in order to understand which computational models I intend to target with Analogical Computational Realism (ACR). Only specific computational models are relevant to realism. First, we should distinguish ordinary computational models from those computational models that count as computational explanations (Piccinini, 2015, pg. 23). A computational model is ordinary if it is used to provide a computational description of some process (cognitive or otherwise) rather than an explanation of what the system does. For example, we can use computers to simulate tornadoes to make sense of specific features (such as invisible swirling vortices), but computation is not being used to explain the phenomenon. Put differently, the computational model describes the system, but computation is not attributed to the system.

Alternatively, sometimes computational models are used to attribute genuine computations to the target system while maintaining that the behavior of that system is itself the result of computation. This difference is important because computational models within computational neuroscience are used in both ways (even sometimes, the same model is used both ways in different contexts). ACR is meant to apply specifically to cases of the latter when the computational model is used to attribute genuine computation to the target system where the target system is the brain. An example of a computational model that counts as a computational explanation in the sense that ACR would apply includes CORnet-S (Kubilius *et al*., 2018a). CORnet-S is an artificial neural network (ANN) that is used to model object recognition tasks in humans. CORnet-S uses a family of architectures that are meant to serve as conceptual analogs to visual areas V1,

V2, V4, and IT, along with a linear decoder that maps from the population of neurons in the model's last visual area to its behavioral choices (Kubilius *et al*., 2018a, pg. 3).

3 HOW TO BE A REALIST ABOUT COMPUTATIONAL NEUROSCIENCE

In this section, I will describe the role that a theory of physical computation plays in underpinning ACR, followed by a specification of the view itself.

3.1 The nature of physical computation

There are different ways to address questions about how computational descriptions relate to the brain. For example, some argue for an account that specifies the implementation relation, the relationship between a formal computational structure and a physical system (Chalmers 1995, 1996; Godfrey-Smith, 2009b; Klein, 2008; Chrisley, 1994; Millhouse, 2019; Scheutz, 2001, Anderson & Piccinini, 2024). These are our mapping views. This type of view is the type that Chirimuuta criticizes. Others offer what can be called a "theory of interpretation," which provides an account of how to understand which computational process a system performs (Pylyshyn, 1986; Dietrich, 1989; Sprevak, 2010, 2018; Shagrir, 2006, 2022; Rescorla, 2013).[15] This family of views includes what are often called "semantic theories." Finally, some have focused on characterizing how to understand what it means to be a mechanism with the function of computing (Piccinini 2018; Anderson & Piccinini, 2024, Chapter 9). While these theorists all intend their view to capture computation in physical systems simpliciter, they also intend for their view to apply to the computational cognitive sciences. For example, Shagrir (2022) argues that we can understand which computations the brain performs by looking at how neuroscientists model cognitive processes,[16] and Piccinini extends his mechanistic account to neural computation (Piccinini, 2020). Whether any of these views provides the ultimate way to underpin computationalism is yet to be seen. However, progress is being made on this topic independently of the success of computational modeling in the cognitive sciences.

Theories of computational implementation, as Chirimuuta points out, ask about the relation between a formal computational structure and a physical system. This question applies to neuroscience in

[15] The distinction between theories that address implementation and theories that address interpretation is not always made distinct in this way. I draw this distinction based on previous work. (Williams, 2023).

[16] Shagrir does not explicitly call his view a kind of *perspectival* realism about computational models. However, there are many places where he implicitly seems to adopt a perspectival realism as I read him. This is apparent when we consider his claim that what a system computes is a matter of *context*. This is not uncommon in semantic accounts. However, when it comes to neural computation the context becomes something like what the scientists say—so we should adopt the function ascribed to the brain by the scientists based on their diagnosis (Shagrir, 2006; 2022). This strikes me as a type of perspectival realism where we are asked to take the perspective of the scientist doing the modeling and experimentation to identify which computations the brain performs.

that it asks about the relation between a given computational model and the brain. Recall that Chirimuuta argued that implementation views have problems, among them is the fact that they can be trivialized and that they ignore that computational models rely on abstraction and idealization. Setting aside issues with trivialization, I agree that philosophers offering an account of implementation have largely ignored the role of idealization in neuroscience. For example, Shagrir (2022) mentions idealization in his book *The Nature of Physical Computation*, but when he does, he references the idea that the "human computer is an idealized entity," meaning it operates under *ideal* conditions—we idealize when we describe computational processes in the sense that we characterize mental computations in terms of correctness (competence) rather than discussing the times where we get things wrong (performance) (Shagrir, 2022, pg. 42-43). In another place, he very quickly mentions that computational models "involve at least some degree of approximation and idealization" (Shagrir, 2022, pg. 230), and he gestures toward how idealizations in the model do not hold in actual biological systems, but he is silent on to what extent and what impact that has on a theory of physical computation (Shagrir, 2022, pg. 237). The same goes for Piccinini (2015) when he presents his mechanistic account of physical computation. Piccinini (2020) comes closer to describing the role of idealization in modeling, but only insofar as it pertains to mechanistic explanation. In the most recent book on physical computation, Anderson and Piccinini (2024) mention idealization when it comes to computer simulations in several places when discussing pancomputationalism, but they do not go into detail about how computational models in neuroscience rely on idealization to model the brain (Anderson & Piccinini, 2024, pgs. 18-19 and 24-25).

While a theory of physical computation should not be held accountable for this shortcoming, not acknowledging the immense role that abstraction and especially idealization play may lead to the assumption that the success of a theory of physical computation (of whatever type) permits one to think that the brain implements a given computational model or that the model can tell us which computations the brain literally performs. Notice that this is the same problematic inference touched on by Chirimuuta regarding the relationship between computationalism and computational modeling but from the opposite direction. In much the same way that the success of computational modeling should not underpin the computationalism thesis, the success of a theory of physical computation cannot underpin a realism toward computational models. What Analogical Computational Realism does is strike a balance between what computational models can tell us and how an independently supported computationalism thesis can help to underpin a realist position toward computational models.

To show how this is possible, consider that views addressing the nature of physical computation do so in different ways. A theory of computational implementation addresses the relation between a formal computational description and a physical system—often doing so by defining some kind of mapping relation between a formal structure and the causal structure of a physical system. Chirimuuta's point is that

in light of abstraction and idealization, we should not think that a theory of this type should be used to define the relation between a computational model in the brain. Notice, though, that this does not mean that we should reject a theory of implementation when it comes to identifying whether a physical system, in general, implements a computational structure. All it means is that a theory like this cannot help us understand when the *brain* implements a computational *model*. A way to avoid this might be to propose some other theory of implementation that doesn't map formal structures to physical systems in the way that these theories typically do.

However, we do not need to overhaul mapping views. There are other options when it comes to thinking about the nature of physical computation. That is, not all theories that address the nature of physical computation do so by defining the implementation relation. Instead, some accounts (namely semantic theories) argue that computation is necessarily a representational process, so what makes the brain a computational system is that it represents in the correct way. There are reasons to be hesitant about these views if they ascribe semantic content based on a computational model (for the same reasons, we shouldn't think that the brain implements the model). But a view like this does not depend on mapping a computational model to the brain (although it might be compatible with one). Yet another option is to take a mechanistic approach to the nature of physical computation. Such a view does not relate a formal computational structure to the brain, and it does not require semantic interpretation. Instead, it asks about the nature of a computing mechanism irrespective of its relation to a formal computational description (Piccinini, 2015).[17] Such a view relies on characterizing what is required for a mechanism to have the function of computing. Finally, some have been interested in making sense of analog computation, and they propose that the brain is best understood as a kind of analog computer (e.g., Maley, 2018). Such a theory may propose a way for us to understand neural computation in terms of analog computation. Indeed, this is an ongoing debate that is continually evolving, and we should expect further developments.

What this means, though, is that we have different ways of addressing whether the brain should be understood as a computing system. Thus, we can pull issues about whether the brain counts as a computing system apart from issues about how computational models relate to the brain. Thus, the success of a theory of physical computation allows one to endorse a realism toward computational models in that those models help to make sense of the actual computational processes in the brain. While there is still an ongoing debate in the philosophy of computation as to which theory we should adopt, I will use this ongoing research program to take on the assumption that the brain performs computations, and I will use this assumption to support Analogical Computational Realism. I will, therefore, speak *as if* there is a consensus on how we

[17] For an account of how The Mechanistic Account of Physical Computation differs from a theory of computational implementation, see Williams (2023) and Williams (2024).

should understand the nature of physical computation and assume that the computationalism thesis has been sufficiently confirmed.

3.2 Analogical Computational Realism

Analogical Computational Realism (ACR) is a realist view in virtue of being committed to computationalism. However, it does not prescribe that we should take computational models as providing literally true descriptions of the computational architecture of the brain, nor should we think that the computations performed by the model should be ascribed to the brain, i.e., we should not interpret the brain as performing the computations used to define the model. Instead, we should take a relevant computational model to provide a how-possibly explanation about computation in the brain. The "analogical" aspect of the view is understood in terms of a comparison between two ways of explaining a phenomenon. The first is to explain the phenomenon by modeling it using formal tools such as mathematics (the computational model). The second is to explain a phenomenon by investigating the material system responsible for the phenomenon (the brain). The two ends of the analogy are tethered to each other by virtue of the mind-independent fact that the brain literally performs computations.[18]

This view requires bringing together two types of explanations that work together to make sense of a cognitive phenomenon. The first type of explanation focuses on the causal-mechanical features of the brain, what may be called the "ontic" explanation—one given in terms of the relations among features of the world (Craver, 2014). However, I will set this part of the analogy aside, reserved for future work, and I will focus on the other type of explanation. The second type of explanation involves the computational model. A computational model can be understood as providing a how-possibly explanation in the modal sense, where the explanandum of a how-possibly explanation will be whatever modal fact describes the possibility in question (Brainard, 2020). As Brainard notes, the role of simulations in furnishing how-possibly explanations has received relatively little attention from philosophers of science (Brainard, 2020, p. 4). Her account, by her lights, does not focus on computation either. In what follows, I will do some of this work by adapting parts of her view in a way that I think helps to make sense of the special role that computational modeling plays in theorizing about the brain and how these models, in particular, can be seen as providing a how-possibly explanation.

[18] This is not to suggest that models generated based on the material system do not rely on mathematics (or even computation). One way we might see the difference is that we can understand the explanations as residing at different levels in Marr's framework, where the computational model is found at the algorithmic (or representational) level while the material explanation is found at the implementation level. I also remain neutral on whether computational models should be understood as mechanistic explanations, which means that I make space for the view that computational models can serve as genuine explanations even if they are not mechanistic.

Computational models play a special role in theorizing because, unlike some other types of models, they often lack transparency, meaning complex computational systems are often opaque, challenging the ability to give transparent scientific explanations. Creel (2022) describes three types of transparency types when it comes to computational models. The first is functional transparency, or a lack of knowledge about the algorithmic functioning of the system, where to have functional transparency is to know which algorithm the system instantiates. The second is structural transparency, or knowledge of how the algorithm is realized in the code. Because an algorithm can be multiply realized in code, it is possible to know the algorithm that the code realizes but not know *how* the code realizes it. Finally, run transparency involves knowledge of the program as it was run in a particular instance, including the hardware and input data used. These transparency issues lend epistemological support to the idea that we should not take computational models to provide literal descriptions of computations in the brain if only because we do not have access to the internal mediating structures and algorithms of the model and thus, we cannot say whether the brain performs those algorithms or whether the brain shares the same instantiation.

Given these transparency issues, what about the model (rather than its inputs or outputs) provides the how-possibly explanation? To start, and as a surprise to nobody, I am not going to offer an account that solves transparency issues. However, there is an active and growing research program that focuses on increasing transparency in existing computational systems. These efforts have seen some success. For example, as Creel (2022) describes, there has been work on creating an algorithm that generates post-hoc decision explanations. One such algorithm is LIME, an algorithm that reduces the opacity of existing machine learning classifiers (Ribeiro et al., 2016). LIME (Local Interpretable Model-agnostic Explanations) aims to explain the predictions of a classifier by fitting a linear model to the pattern of its prediction given the input data. Such an algorithm increases functional transparency by providing access to the decision space most relevant to the token classification. Programs like this, along with other efforts, provide an optimistic nudge toward the idea that transparency issues might be tackled. I will take this optimistic line and argue that computational models may be capable of providing a how-possibly explanation in some cases.

Before stating what is required of a how-possibly computational explanation, a similar point should be made for the brain as well: we cannot read off the content of neural signals from material brain processes. That is to say, we cannot perceive (even with our best tools) the contents of computational processes in the brain. Put in information-theoretic terms, we cannot, by looking at the material stuff, determine the message carried by the material states: we can perceive neurons, and we can perceive that they are active, but we cannot perceive the content of the message that they are transmitting—we just infer that they are transmitting *some* message and we use our theoretical tools to hypothesize about what that message might be. To restate: We can infer that information is being moved through the brain by tracking it indirectly (e.g.,

tracking oxygen in blood flow, electrical impulses, etc.). But those measurement tools allow us to measure physical quantities and the transference of those quantities through the system; they do not allow us to, even indirectly, read off the messages contained within the material states. To ascribe a computational *process* to the brain is to semantically interpret the computational states of the brain. Sometimes, that interpretation can be given in terms of an algorithmic procedure where the algorithm invokes arithmetical operations that include procedural interpretations (e.g., Rescorla, 2013), other times, the interpretation can be given in terms of a mathematical function (e.g., Shagrir, 2022) where the content of the neural states is understood as representing mathematical objects. Relating algorithms and mathematical functions to the brain is one way of explicating the representational content of a computational process.[19]

Additionally, because tracking the message that is being transferred through the brain helps to define the computational structure—or path that information takes through the brain—identifying the computational structure of the brain is also opaque to us. Appreciating this helps to demonstrate one of the problems with taking computational models as literal descriptions of the computational processes in the brain. First, the models themselves are opaque to us, and second, the brain is computationally opaque to us in similar ways (which is one reason we use computational models in the first place!). The difference is that we can probe computational models to help address the transparency issues, whereas we cannot do the same with the brain (if we could, we would not need computational models).[20] Thus, the model once made more transparent, can serve as a bonafide theory or hypothesis about computational processes in the brain; it can serve as a how-possibly explanation about what computational processes are actually performed by the brain.

To understand what is required of a how-possibly explanation when it comes to computational models, I adapt two of Brainard's (2020) conditions:

1. How-possibly explanations take as their explananda modal facts of the form "it is possible that *p*" (given a set of contextually determined relevant background beliefs).

[19] As opposed to describing the content of computational states in terms of the distal environment. While we cannot observe that type of content either, recognizing that there are different ways to cash out the content of a computational process is important for understanding the similarities between computational processes and how the brain is described using computational tools. That is not to say that computational models are never ascribed contents based on the distal environment, but they are mathematical constructs that are defined in terms of mathematical features such as algorithms and mathematical functions, and both of those features are also ascribed to the brain. This point is subtle and is often underappreciated outside of the literature on physical computation, but these computational features are often understood as content or as representational content (or semantic content). Thus, it is worth appreciating this nuance when it comes to understanding the relationship between computational models and the brain and how they are used within computational neuroscience to describe and understand neural processes.

[20] This point also helps make sense of why there are two parts to the analogy: the computational model and the material system. The idea is that neither can give the complete story about computation in the brain. Instead, they work together to provide an understanding of computation in the brain, tethered together by the thesis that the brain counts as a computing system to begin with.

2. How-possibly explanations must be potential why-actually explanations in the following sense: if their content were true, they would succeed at explaining why the state of affairs in question actually obtained.

To sum, what makes the computational model a how-possibly explanation is that it describes a possible state of affairs, and the explanation is a reasonable contender for (eventually) explaining how that state of affairs obtains.

Because the computational model should not be taken literally or taken to be implemented in the brain, we can understand the model as providing a computational "real pattern" in the Potochnik (2017) sense. Potochnik proposes that real patterns are the target of scientific models. What makes a pattern real is that it depicts regularities in a phenomenon (in the Dennett sense). Thus, we can understand computational models as depicting (computational) regularities in the (computational) brain. Potochnik conceives of regularities in the phenomena as being real because they track causal patterns. However, guided by the how-possibly framework, I take the regularities to track *possible* causal patterns and to help us understand *possible* semantic interpretations of those patterns. Put differently, the patterns tracked by the model provide a *hypothesis* about the computational features of the brain. Not all computational models will fill this role, though, because not all computational models are meant to provide explanations of this sort. However, some are, and often those models are developed carefully in such a way that focuses on good experimental design, are made consistent with values and parameters that best describe the behavioral data (and that those parameters are recoverable), and validated among other steps (Wilson & Collins, 2019). How-possibly computational explanations serve as a hypothesis about a computational process in the brain.

3.4 ACR and Computational Neuroscience

ACR is a realism tailored for computational models that is meant to inform philosophers interested in interpreting results from computational neuroscience. What ACR does is allow one to be a realist about certain computational models by taking the model to provide a how-possibly explanation in the form of a hypothesis about computation in the brain without rejecting that the brain performs computations and without being committed to the idea that the brain implements the model. Also, by requiring the establishment of the computationalism thesis independently, it allows the philosopher to make claims about computational processes without simply *assuming* computationalism. An upshot of taking on this view is that it maps onto the way that neuroscientist typically characterize the role that their models play in a theory. It is no secret that neuroscientists know that their models are not accurate depictions of brain processes (even if they sometimes make claims that seem to contradict that knowledge). For example, Driscoll *et al*. (2024) use a Recurrent Neural Network (RNN) to try and understand cognitive flexibility, such as rapid learning and task switching. They propose that they have "identified an algorithmic neural substrate for

modular computation through the study of multitasking artificial recurrent neural networks" (Driscoll *et al*., 2024, pg. 1349). If we were to take this claim literally, we might think that the RNN tells us something explicit (i.e., *something true*) about the actual implementing mechanisms in the brain responsible for those specific cognitive flexibility tasks. However, we should notice that the neuroscientists acknowledge that the model serves as a *hypothesis* based on a *simplified prox*y for biological neural networks:

> Our results are based on artificial systems, lacking the complexities of real brains. We used simplified networks without diverse cell types or prescribed architectures and only applied noisy static inputs. Although our learning rules are not biological, we hypothesized that optimized artificial neural networks and the principles that we uncover from them are informative about biological neural circuits based on principles of optimality and robustness (Driscoll *et al*., 2024, pg. 1361).[21]

To see the difference between ACR and a version of realism that takes a literal interpretation of the model, consider the above example again. The modelers state that the RNN was defined by a set of mathematical functions that it uses to calculate the input vectors to produce an output. What this means is that to produce the output from the input, the variables in the input vectors are calculated according to those functions to produce the output (either a 1 or no response). A literal interpretation of this is that the brain performs those very functions when it solves the same task. Put differently and using the language from the experiment, we would say that when the brain performs the 'ReactCategoryPro' task, it does so by performing the function that defines the RNN. We see this type of literal interpretation in Shagrir (2006, 2022) when he says that "the methodological role that input-output modeling plays in computational theories… helps to reveal the mathematical input-output function that the system computes" – the system he has in mind in this statement is the brain (Shagrir, 2022, pg. 249). He takes the function that defines the model as "the" function that the brain solves during that same task.[22]

Alternatively, ACR advises that philosophers refrain from taking the function that defines the model to be the function that the brain computes during the task. To give a concrete application of ACR, consider Driscoll et al. again, but now, by looking at what they say about the structure of the computational model. The ACR approach toward their results is that the RNN provides a highly simplified model of how

[21] I want to pause here and say that I am not trying to interpret the mental states of neuroscientists or even make claims about their commitments. It is common knowledge among philosophers of neuroscience that it is quite difficult to interpret what neuroscientists *mean* when they say certain things. This is a challenge especially because it will vary from scientist to scientists. All that I mean to show here is that scientists are *aware* that their models rely on abstraction and idealization and that they are providing a representation of the brain that is not complete nor is it fully accurate.

[22] As I mentioned in footnote 17, Shagrir seems to offer a kind of perspectival realism that is difficult to reconcile with the claim that the scientific context *literally* determines the function computed by the brain. So, I characterize his view here by following the way he describes his view combined with his metaphysical stance that computation is necessarily a representational process, but I leave open whether he means to attribute the function in this strongest sense. However, the example still serves its illustrative purpose even if Shagrir is not ultimately committed to this exact view.

it might be *possible* for the brain to execute certain cognitive tasks, such as rapid task switching, by re-using the same neural structures when it comes to similar tasks that are composed of the same compositional building blocks. The model provides a how-possibly story about how certain structures in the brain may serve as basic building blocks for the same task (cases of neural reuse for different tasks) by showing how the computational process can be understood as relying on an attractor-ring structure (as described by the authors). But we should not go further and say that the brain implements the attractor ring. Put differently, the attractor ring serves as a hypothesis for the computational pattern of behavior in the brain responsible for how the task is executed—it should be understood as a how-possibly explanation.

4 TAMING ANALOGICAL COMPUTATIONAL REALISM

Notice that ACR may direct us to be a realist about many different computational models as there are many different ways to computationally model the same process. This makes ACR quite liberal. While I think that pluralism is a good result, it can and should be tamed for fear of being *too* liberal. Pluralism about models typically starts from the assumption that models deliver partial, interest-dependent, and contingent representations of the world (Ludwig & Ruphy, 2021). In some cases, developing a plurality of partial models of a target system constitutes a step toward a richer, integrative representation of a system, especially considering that some modelers may wish to track different patterns within the system (Potochnik, 2017). This approach is particularly relevant to neuroscience, where the brain is modeled at different grains, from different perspectives, and with different explanatory aims. Thus, allowing for model pluralism is a good result for ACR because it allows us to approach understanding the brain from many directions.

However, when it comes to model pluralism, it is possible to computationally model the same process at the same grain in different ways. So, what we end up with are multiple, often quite different, and sometimes inconsistent models.[23] In some cases, though, it may be possible to choose between the different models to tame the pluralism. Taming ACR is the idea that we can narrow down the candidate hypotheses if we look at certain ways in which neuroscientists are already working on this issue. In this section, I will propose two ways we may go about doing this. The first involves model comparison techniques, while the second considers how certain benchmarks can be used for comparing and choosing between computational models. What I aim to do in this section is give two examples of how it is possible to narrow down candidate models by focusing on ways in which neuroscientists have been engaging in the project, thus keeping the strategy consistent with actual neuroscience practice.

[23] Chakravartty (2010) discusses the issue of model inconsistency and its relation to realism.

4.1 Model Comparison Techniques

Golan *et al*. propose a framework that adjudicates between theories by pitting neural networks against each other to efficiently compare theory predictions of neuronal and behavioral responses. They deploy this technique using deep neural networks that classify images:

> Even when there is a considerable difference in test accuracy between two models, the more accurate model is not necessarily more human-like in the features that its decisions are based on. The more accurate model might use discriminative features not used by human observers. DNNs may learn to exploit discriminative features that are completely invisible to human observers (4, 5). For example, consider a DNN that learns to exploit camera-related artifacts to distinguish between pets and wild animals. Pets are likely to have been photographed by their owners with cellphone cameras and wild animals by photographers with professional cameras. A DNN that picked up on camera-related features might be similar to humans in its classification behavior on the training distribution (i.e., highly accurate), despite being dissimilar in its mechanism. Another model that does not exploit such features might have lower accuracy, despite being more similar to humans in its mechanism. To reveal the distinct mechanisms, we need to move beyond the training distribution (Golan *et al.,* pg. 29330).

The concern is that models that produce behavior that is most similar to humans may be doing so in a way that diverges from how humans perform the task. So, better performance does not always mean that the model best captures how humans might be performing the task. A way to address this is to pit models against each other and subject them to repeated testing using controversial stimuli (a sensory input that elicits clearly distinct responses among two or more models) and compare their behavior to data collected from human responses.

In their first experiment, the authors pitted models that were trained using the MNIST data set (a large training database of handwritten digits) against each other. Nine models covering five computational categories were included: discriminative feedforward models, discriminative recurrent models, adversarially trained discriminative models, a reconstruction-based readout of the Capsule Network, and class-conditional generative models. In the second experiment, the authors pitted models that used the CIFAR-10 data set (a dataset consisting of 60,000 color images in 10 classes with 6,000 images per class). Seven models that fell into five model families (largely overlapping with the model families tested in experiment one) were used: discriminative feedforward models, a discriminative recurrent model, adversarially trained discriminative models, a class-conditions generative model, and a hybrid discriminative-generative model. The results of their experiments were that generative models may better capture human object recognition, but they are careful to note that none of them are functionally *equivalent* to the process that generated the human responses (Golan *et al.,* pg. 29336).

Something important to keep in mind when considering model comparison strategies such as the ones described above is that ACR prescribes being a realist about models that provide computational explanations of brain processes. While researchers sometimes make claims that models like the ones above are models of human cognition, they aren't always used this way. We should be mindful of this difference because sometimes models in artificial intelligence are designed to perform cognitive tasks, but we should not, therefore, think that they are models of human cognition. Keeping this difference in mind, we can still see that this type of strategy may be fruitful when it comes to adjudicating between models in neuroscience.

4.2 Criteria-based model choice

Returning to the CORnet-S example, we might consider being a realist about particular models that meet specific benchmarks. For example, CORnet-S is meant to be a model of a biological visual system. Kubilius *et al*. are interested in models that consider what we know about the brain from neurobiology. They argue that many of the artificial neural networks (ANN) in use have lost sight of "the connection to neurobiology [and have] grown far more murky in that it is unclear which, if any, model layer(s) are putative models of specific ventral stream cortical areas*" (*Kubilius *et al*., 2018a, pg. 1). I will focus on the general criteria that they give which includes the following features: predictive, compact, and computable (Kubilius, 2018b).

The 'predictability' criterion asks whether the model predicts unseen data. This criterion is motivated by the idea that to understand a phenomenon we must be able to predict all of the explainable variance in the data for any input in the domain over which the model is claimed to hold (Kubilius, 2018b, pg. 110). To incorporate the predictability criterion, Kubilius *et al.* (2018a) use 'Brain-Score' to quantify whether their model meets the requirement. Brain-Score is a composite of multiple neural and behavioral benchmarks that score any ANN on how similar it is to the brain's mechanisms for core object recognition (Schrimpf et al., 2020). The 'compact' criterion involves looking at how many principles the model depends on. This criterion looks for simple procedures used to solve complex models (not necessarily that the model itself must be the most simple) (Kubilius, 2018b, pg. 111). When developing CORnet-S, the authors preferred (from the ones suggested by Brain-Score) the simpler models that were easier to understand and more efficient to experiment with. Finally, 'computability' is the criterion that for any model, we should know what counts as an input and how to compute the input to produce the output. What this means is that the solution to the problem should be implementable in the hardware (Kubilius, 2018b, pg. 111). To meet this requirement, Kubilius *et al.* (2018a) preferred a model that would act like a participant in an experiment by receiving the same instructions and producing outputs without any free parameters left for a researcher to fine-tune (Kubilius *et al.*, 2018a, pg. 2).

While Kubilius *et al.* (2018a) used a criteria-based model selection that narrowed models down based on whether they were more brain-like, it is not claimed by the authors that the model depicts how the

actual visual system works. For example, in the discussion section of the paper, the authors acknowledge how the model is detail-poor in the sense described in section 2:

> While our long term goal is a model of all the mechanisms of the ventral stream, we do not claim that CORnet models are precisely biomimetic. For example, we ignored that visual processing involves the retina and lateral geniculate nucleus (LGN)…" (Kubilius, *et al*., 2018a, pgs. 7-8).

And also how it involves the use of idealized features:

> …Most importantly, making sure the model's nonarchitectural parameters can be set by gradient descent (i.e., the model can be trained) and that the model occupies few GPUs during train time. So, for instance, adding a skip connection was not informed by cortical circuit properties but rather proposed…as a means to alleviate the degradation problem in very deep architectures (where stacking more layers results in decreased performance" (Kubilius *et al*., 2018a, pg. 8).

CORnet is a paradigm example of a computational explanation in the sense that it attributes computation to the system it describes, but the use of both abstraction and idealization is explicitly acknowledged. This is an indication that relying on specific strategies to narrow down potential models does not deny that even the narrowed candidate hypotheses rely heavily on abstraction and idealization—the crucial motivator for ACR.

In this section, I have described two ways we might go about narrowing down the computational models that serve as hypotheses. I take a main tenant of realism to be the idea that we don't want to be a realist about every model that scientists propose. Instead, there must be something about the model that warrants realism. Pointing out these considerations when it comes to realism about computational models puts some of the responsibility onto the philosophers when it comes to theorizing about or drawing from computational neuroscience. To do justice to the science and to philosophical theory, we should understand the nature of computational models, what they can actually tell us, and how they contribute to theory development in neuroscience. This includes doing additional research about the type of model being proposed, what alternatives there may be, how that model fits into various frameworks within neuroscience, and what motivations drive the use of a specific model.

5 CONCLUSION

In this paper, I have provided a way to be a realist about the computational models in neuroscience. This realism proposes that we take a subset of computational models to provide a how-possibly explanation of the computational features of the brain in the sense that the model should be understood as a hypothesis regarding the computational patterns in the brain. The "analogical" part of the view is understood as a comparison between two different ways of explaining a phenomenon. The first is to explain the

phenomenon by modeling it using a computational model. The second is to explain a phenomenon by investigating the material system responsible for the phenomenon. The two ends of the analogy are tethered by the computationalism thesis, which is independently supported by a theory of physical computation. Computational models are understood as depicting possible causal patterns in the brain, which are understood as real patterns—as a hypothesis about computational processes in the brain.

Additionally, I have suggested two ways that we might go about taming ACR by reducing the number of models that we are a realist about. We can tame ACR in different ways—I have offered only two. Importantly, ACR is a view that takes seriously the idea that some computational models genuinely can provide insight into the computational processes in the brain. It is too soon to say which strategies will be the best ways to narrow down all possible models, just as it is too soon to say which models provide the best how-possibly explanations such that they could reasonably lead to a how-actually story. Computational neuroscience is still a relatively new field, and the computational tools being used are being developed in real-time. But despite the currently evolving nature of the field, we can still make progress on thinking about what computational models can help us understand about computation in the brain and how we should use computational theories in philosophical inquiry into the computational nature of cognition.

Acknowledgments I wish to thank Daniel Burnston, J. Brendan Ritchie, David Barack, Victor Verdejo, Andrew Rubner, Zoe Drayson, members of the audience at the 2024 meeting of the Society for Philosophy and Psychology, the postdoc working group in the Center for Humanities at Washington University in St. Louis (especially Claudia Carroll) for comments on the various drafts of this paper, and finally, several anonymous reviewers whose careful and thoughtful comments helped to refine and improve the paper.

References

Anderson, N. and Piccinini, G. (2024). *The Physical Signatures of Computation: A Robust Mapping Account*. Oxford University Press.

Bechtel, W., Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton University Press.

Brainard, L. (2020). How to explain how-plausibly. *Philosophers Imprint.* Vol. 20, 13.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Oxford: Oxford University Press.

Chakravartty, A. (2010). Perspectivalism, Inconsistent Models, and Contrastive Explanation. *Studies in History and Philosophy of Science Part A*, vol. 41, no. 4, 2010, pp. 405–412.

Chalmers, D. (1995). On Implementing a Computation. *Minds and Machines*, 4, 391-402.

Chalmers, D. (1996). Does a Rock Implement Every Finite-State Automaton? *Synthese*, 108, 310-333.

Chirimuuta, M. (2016). Vision, perspectivalism, and haptic realism. *Philosophy of Science*, vol. 83, no. 5, pp. 746–756.
Chirimuuta, M. (2023). Haptic realism for neuroscience. *Synthese*, *202*(3).
Chirimuuta, M. (2024a). *The Brain Abstracted*. MIT Press.
Chrisley, R. L. (1994). Why Everything Doesn't Realize Every Computation. *Minds and Machines* 4: 403-430.
Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. The MIT Press.
Craver, C. F. (2014). The Ontic Account of Scientific Explanation in In Marie I. Kaiser, Oliver R. Scholz, Daniel Plenge & Andreas Hüttemann (eds.), *Explanation in the special science: The case of biology and history*. Dordrecht: Springer. pp. 27-52.
Creel, K.A. (2022). Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4), pp. 568-589.
Dennett, D. (1991). Real Patterns. *The Journal of Philosophy*. 88(1): 27-51.
Dietrich, E. (1989). Semantics and the Computational Paradigm in Cognitive Psychology. *Synthese* 79: 119-141.
Driscoll, L., Shenoy, K., Sussillo, D. (2024). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nature Neuroscience*. Vol. 27. pp. 1349-1363.
Dupre, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.
Giere, R. (2006). *Scientific Perspectivalism*. The University of Chicago Press.
Godfrey-Smith, P. (2009a). Abstractions, idealizations, and evolutionary biology. In *Springer eBooks* (pp. 47–56).
Godfrey-Smith, P. (2009b). Triviality Arguments Against Functionalism. *Philosophical Studies* 145 (2): 273-295.
Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(47), 29330–29337.
Hacking, Ian, (1982). Experimentation and Scientific Realism, *Philosophical Topics*, 13(1): 71–87.
Jones, M. (2005). *Idealization and abstraction: A framework*. In M. R. Jones & N. Cartwright (Eds.), Idealization XII: Correcting the model. Idealization and abstraction in the sciences (Vol. 86, pp. 173–217). Rodopi.
Kitcher, P. (1993). *The Advancement of Science: Science Without Legend, Objectivity without Illusions*. Oxford University Press.
Klein, C. (2008). Dispositional Implementation Solves the Superfluous Structure Problem. *Synthese* 165 (1): 141-153.
Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.
Kubilius, J. (2018b). Predict, then simplify. *Neurolmage*. Vol. 180. pp. 110-111.
Laymon, R. (1995). *Idealizations and the testing of theories by experimentation*. In P. Achinstein & O. Hannaway (Eds.), Observation, experiment, and hypothesis in modern physical science. MIT Press.
Levy, A. (2021). Idealization and abstraction: refining the distinction. *Synthese*, *198*(S24), 5855–5872.
Ludwig, D, and Ruphy, S. (2021). Scientific Pluralism. *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/win2021/entries/scientific-pluralism
Maley, C. (2018). Toward Analog Neural Computation. *Minds and Machines*. 28(1):77-91.
Massimi, M. (2004). Non-Defensible Middle Ground for Experimental Realism: Why We are Justified to Believe in Colored Quarks, *Philosophy of Science*, 71(1): 36–60. doi:10.1086/381412
Massimi, M. (2023). Epistemic communities and their situated practices: Perspectival realism—a primer, *Annals of the New York Academy of Sciences,* Vol. 1523 (1): 5-10.
McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science*, XVI. pp. 247–273.

Millhouse, T. (2019). A Simplicity Criterion for Physical Computation. *British Journal for Philosophy of Science,* 70: 153-178.
Mitchell, S. D. (2012). *Unsimple Truths: Science, Complexity, and Policy*. Chicago: University of Chicago Press.
Nowak, L. (1992). The idealizational approach to science: A survey. In J. Brzezi´nski & L. Nowak (Eds.), Idealization III: Approximation and truth. Rodopi.
Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press.
Piccinini, G. (2020). *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford University Press.
Piccinini, G. and Maley, C. (2021). Computation in Physical Systems. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/sum2021/entries/computation-physical systems.
Potochnik, A. (2017). *Idealization and the aims of science*. The University of Chicago Press.
Pouget, A., Sejnowski, T.J. (2001). Lesioning a basic function model of spatial representations in the parietal cortex: comparison with hemineglect. *Psychological Review*. 108(3): 653-673.
Psillos, S. (1999). *Scientific Realism: How science tracks truth*. http://ci.nii.ac.jp/ncid/BA44318430
Putnam, H. (1975). *Mathematics, Matter and Method*, Cambridge University Press.
Putnam, H. (1988). *Representation and Reality*. The MIT Press.
Pylyshyn, Z. W. (1986). *Computation and Cognition: Toward a Foundation for Cognitive Science*. The MIT Press.
Rescorla, M. (2013). Against Structuralist Theories of Computational Implementation. *The British Journal for the Philosophy of* Science. 64: 681-707.
Ribeiro, M. T., Singh, S., Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data
Scheutz, M. (2001). Computational versus Causal Complexity. *Minds and Machines.* 11: 543-566.
Sejnowski, T. (2015). Computational Neuroscience in the *International Encyclopedia of the Social & Behavioral Sciences: Computational Neuroscience* edited by J. D. Wright. 2nd ed. Elsevier.
Shagrir, O. (2006). Why We View the Brain as a Computer. *Synthese* 153(3): 393-416.
Shagrir, O. (2022). *The Nature of Physical Computation*. Oxford University Press.
Schrimpf, M, Kubilius, J., Hong, H., Najib, M.J., Rajalingham, R., Elias, I.B, Kar, K., Bashivan, P., Prescott, J., Geiger, R.F., Schmidt, K., Yamins, D.L.K., DiCarlo, J. (2020). bioRxiv 407007.
Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science.* 41: 260-270.
Sprevak, M. (2018). Triviality arguments about computational implementation. In *Routledge Handbook of the Computational Mind*, by Mark Sprevak and M. Colombo, pp. 175-191. Routledge.
Teller, P. (2001). Twilight of the Perfect Model Model. *Erkenntnis*, 55, 393-415.
van Fraassen, Bas C., (1980). *The Scientific Image*, Oxford: Oxford University Press.
Weisberg, M. K. (2007). Three kinds of idealization. *The Journal of Philosophy*, *104*(12), 639–659.
Weisberg, M. (2013). *Simulation and similarity*. New York: Oxford University Press.
Williams, D. J. (2023). *Implementation and interpretation: A unified account of physical computation*. Available from ProQuest Dissertations & Theses Global. (2865991732).
Williams, D. J. (*forthcoming*). Two senses of medium independence. *Mind & Language*.
Williams, D.J. (2024). It takes two to make a view go right. *The Brains Blog*.
Wilson, R., Collins, A.G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*.
Wimsatt, W. C. (1987). (In Press) Re-engineering philosophy for limited beings: Piecewise approximations and reality. Cambridge, MA: Harvard University Press. Originally Published in

Neutral Models in Biology: False Models as Means to Truer Theories (M. Nitecki & A. Hoffman, Eds.). Oxford University Press.
Wimsatt, W. C. (2007). *Re-Engineering Philosophy for limited beings*. Harvard University Press.