Posted on 29 May 2024 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.171701343.33979173/v1 — This is a preprint and has not been peer-reviewed. Data may be preliminary.

How do Computational Models in the Cognitive and Brain Sciences Explain?

Cédric Brun[1], Jan Pieter Konsman[2], and Thomas Polger[3]

[1]Université de Bordeaux École Doctorale Montaigne Humanités
[2]Immuno ConcEpT
[3]University of Cincinnati

May 29, 2024

Abstract

The nature of explanation is an important area of inquiry in philosophy of science. Consensus has been that explanation in the cognitive and brain sciences is typically a special case of causal explanation, specifically, mechanistic explanation (Craver 2007). But recently there has been increased attention to computational explanation in the brain sciences, and to whether that can be understood as a variety of mechanistic explanation. After laying out the stakes for a proper understanding of scientific explanation, we consider the status of computational explanation in the brain sciences by comparing the mechanistic proposal to computational accounts advanced by Piccinini (2015), Milkowski (2013)), Cao (2019), Chirimuuta (2014, 2018), and Ross (Ross 2015, 2023). We argue that many of these accounts of computational explanation in neuroscience can satisfy the same explanatory criteria as causal explanations, but not all. This has implications for interpretation of those computational explanations that satisfy different criteria.

How do Computational Models in the Cognitive and Brain Sciences Explain?

Cédric Brun

University of Bordeaux-Montaigne

Jan Pieter Konsman

University of Bordeaux and CNRS

Thomas Polger †

University of Cincinnati

†Corresponding author: Thomas Polger, thomas.polger@uc.edu

All authors contributed equally to conceptualization, drafting and editing of this project. Author names appear alphabetically.

Brun, Cédric

proxy

Conceptualization (Equal)

Writing – original draft (Equal)

Writing – review & editing (Equal)

Konsman, Jan Pieter

proxy

Conceptualization (Equal)

Writing – original draft (Equal)

Writing – review & editing (Equal)

Polger, Thomas

self

Conceptualization (Equal)

Writing – original draft (Equal)

Writing – review & editing (Equal)

Keywords: computational explanation; computational neuroscience; mechanism; philosophy of neuroscience; methodology

Running Head: Computational Models in Neuroscience

Word count: 8847

Figure count: 0

AUTHOR STATEMENTS

Author Contributions: All authors contributed equally to conceptualization, drafting and editing of this project. Author names appear alphabetically.

Research Ethics: This is a theoretical study; no experiments were conducted

Animal Care: This is a theoretical study; no experiments were conducted

Data Sharing: This is a theoretical study; no data were produced

LIST OF ABBREVIATIONS

3M	Model-to-Mechanism Mapping
HH	Hodgkin-Huxley Model
MDB	More Details are Better
V1	Primary visual cortex

How do Computational Models in the Cognitive and Brain Sciences Explain?[*]

Abstract

The nature of explanation is an important area of inquiry in philosophy of science. Consensus has been that explanation in the cognitive and brain sciences is typically a special case of causal explanation, specifically, mechanistic explanation (Craver 2007). But recently there has been increased attention to computational explanation in the brain sciences, and to whether that can be understood as a variety of mechanistic explanation. After laying out the stakes for a proper understanding of scientific explanation, we consider the status of computational explanation in the brain sciences by comparing the mechanistic proposal to computational accounts advanced by Piccinini (2015), Milkowski (2013), Cao (2019), Chirimuuta (2014, 2018), and Ross (2015, 2023). We argue that many of these accounts of computational explanation in neuroscience can satisfy the same explanatory criteria as causal explanations, but not all. This has implications for interpretation of those computational explanations that satisfy different criteria.

Introduction

Recent years have seen a resurgence of work on perhaps the most central problem in philosophy of science: the nature of scientific explanation. Lately, attention has focused specifically on computational explanation in the cognitive and brain sciences. One

[*] Author names appear alphabetically. All authors contributed equally to this project.
ACKNOWLEDGEMENTS REMOVED FOR ANONYMOUS REVIEW.

reason is that computational explanation has become much more prevalent in the cognitive and brain sciences since the last period of philosophical focus on the nature of explanation (cf. (Salmon and Fagot-Largeault, 1989). A second reason is that the general shift toward mechanistic strategies for explanation in the sciences (e.g., Bechtel and Richardson, 1993, Machamer et al., 2000), particularly as applied to the brain sciences, has highlighted the lack of any adequate account of computational explanation in these fields. In this paper, we focus on the question of how computational explanations work in the cognitive and brain sciences, with attention to their implications for the practices of the cognitive and neurosciences.[1]

Cognitive science was born in the 1950s at a time when some underlying disciplines like psychology and anthropology were themselves undergoing transformation — and other sciences, such as computer science and neuroscience, still had to fully emerge (Miller, 2003). More than half a century later, the marriage between cognitive science and neuroscience in the form of cognitive neuroscience has encountered both enthusiasm and criticism. Enthusiastic authors seem to agree that cognitive neuroscience became possible when imaging techniques at least allowed the study of brain activity in human subjects performing cognitive tasks (Kosslyn and Shin, 1992; Kriegeskorte and Douglas, 2018; Pereira, 2007). Critics, however, have deemed this marriage "troubled" in view of the differences in concepts and methods between

[1] The term *brain sciences* is sometimes preferred to *neuroscience*. Cognitive science, at least early on, was accompanied by brain sciences rather than by neuroscience. Indeed, the term *neuroscience* was only put forward by the United States National Academies Committee on Brain Sciences at the end of the 1960s (Altimus et al., 2020). In addition, the term *brain sciences* allows one to emphasize other levels of nervous system organization rather than giving priority to neurons. It is natural for us to consider neuroscience to be part of the brain sciences. However, there is also a generic use of the term *neuroscience* that covers both the cognitive and brain sciences. We use these terms as generally as possible.

cognitive science and neuroscience (Cooper and Shallice, 2010). For example, cognitive and brain sciences, including neuroscience, frequently differ over how they approach and evaluate explanations. It is familiar that neurobiology often favors mechanistic explanations and cognitive sciences often favor computational explanations.

Mechanistic explanations have dominated the attention of philosophers of neuroscience at least since Carl Craver's groundbreaking book, *Explaining the Brai*n (2007). But the first word is not the last word, and the details of mechanistic models are a matter of great dispute. Glennan & Illari propose a minimal definition, according to which "[a] mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon" (Glennan and Illari, 2018: 2). Our aim here is not to defend an account of mechanisms, but rather to relate how philosophers of science have compared mechanistic explanations to other accounts when it comes to the cognitive and brain sciences.

For example, Craver argues that "[n]ot all models are explanatory," but that "[m]echanistic models are explanatory" (Craver, 2006, p. 355). According to him, other models, such as the Hodgkin-Huxley model describing the changes in permeability of a neuron to sodium and potassium in equations, "are data summaries", or "phenomenal models" (Craver, 2006). This stance makes sense in the light of his vision that "to explain something … is to show how it fits into the causal structure of the world" (Craver, 2009, p. 578). Building on the basic mechanistic framework, William Bechtel argues for mental mechanisms as "mechanisms that process information" (Bechtel 2008). He further argues that accounts of mechanistic explanation should not only look "down" to determine in more detail the entities or parts and activities or operations that

compose a mechanism, but also ought to look "up" as "to situate a mechanism in its context" (Bechtel 2009).

Bechtel and Oron Shagrir have interpreted David Marr's famous levels of analysis (viz., computational, algorithmic or representational and implementational) as perspectives to better understand information-processing mechanism perspective. Accordingly, "[t]he computational perspective provides an understanding of how a mechanism functions in broader environments that determines the computations it needs to perform," the "algorithmic perspective offers an understanding of how information about the environment is encoded within the mechanism and what are the patterns of organization that enable the parts of the mechanism to produce the phenomenon" and "[t]he implementation perspective yields an understanding of the neural details of the mechanism and how they constrain function and algorithms" (Bechtel and Shagrir 2015: 312).

Emphasis on so-called levels in the questions and positions just summarized may give the false impression that things are or can be neatly separated. While the authors who share such concerns often seem to favor a pluralism of explanatory styles (Krakauer et al., 2017; Potochnik and Sanches de Oliveira, 2020), questions remain as to what kinds of explanations are required in order for cognitive sciences to explain and how computational or mathematical explanations and models should be considered.

Kaplan and Craver, for example, argue that "dynamical and mathematical models in systems and cognitive neuroscience explain (rather than redescribe) a phenomenon only if there is a plausible mapping between elements in the model and elements in the mechanism for the phenomenon" (Kaplan and Craver, 2011). In a

similar spirit, Piccinini and Boone have suggested "a framework of multilevel neurocognitive mechanisms that incorporates representation and computation" (Boone and Piccinini, 2016). For these authors, the explanatory value of mathematical and computation models in cognitive science seems to be conditioned on the explanatory value of underlying mechanisms. This preference makes sense in light of the close connection between mechanistic and causal explanation, to which we return, below.

Can all computational explanations and models be treated as mechanistic explanations and models? And, why should we care? In section 2, we review the relevance of theories of explanation for the cognitive and brain sciences. In section 3, we examine the importance of understanding explanation in the special case of causal or mechanistic explanations. This positions us, in section 4, to raise the question of how computational models and explanations relate to those causal or mechanistic models. If computational models just are mechanistic models, then we know how the connection will go. But if they are not, then there is more work to be done. We conclude by outlining that work, in section 5.

Why care about explanation at all?

Before we dive into the topic of computational explanation, it is worthwhile to consider why discussion of explanation is profitable. Philosophers may suppose that it is obvious that studying and theorizing about explanation is a worthwhile endeavor; but this may be less the case for (neuro)scientists who are often trained in applying one type of explanation.

To those who do not already think that the study of explanation is intrinsically valuable, or those who doubt that its intrinsic value is sufficient reason for its pursuit, several things can be said. First, it is plausible that an understanding of explanation itself is necessary if we are to give any principled accounts of when and how explanations succeed or fail. If it is right that the height of a flagpole explains the length of the shadow it casts but the length of the shadow does not explain the height (Bromberger, 1966), then we would like to know what makes the difference. If the fact that Josephina uses contraception explains why she does not become pregnant but Joseph's use of contraception does not explain why he does not become pregnant (Salmon, 1971), then we would like to know what makes the difference. And if the number of strawberries in the container being 21 explains why it cannot be evenly divided into two equinumerous containers without cutting strawberries but does not explain why it cannot be divided into two containers of equal weight (Lange, 2013), then we would like to know the difference.

The claim here is not that a theory of explanation should be in any way prior to the conduct of the various sciences. On the contrary, it is familiar within the sciences to raise questions about the distinction between description and explanation (e.g., with respect to the status of either law-based or dynamical systems explanations), when explanation is finished or complete (e.g., with respect to whether causal or reductive explanations must terminate at some point), or what kind of empirical support is required for adequate explanation (e.g., with respect to the "replication crisis," "p-hacking," or whether reductive explanations must entail a priori principles that enable transcription of entities from one higher-level theory to those of a lower-level theory

(Casadevall and Fang, 2008; Claxton et al., 2005; Hanna, 1969; MacQueen, 2013; Reese, 1999)). So these kinds of concerns are not specifically "philosophical" or "a priori" in any distinctive way.

Relatedly, there will be important questions about the relations between various explanations or candidate explanations. This will even be the case if there is only one kind of explanation — for then we will need to know when explanations compete with one another and, if they do, how to choose among them. Yet it is plausible that there is more than one kind of explanation or explanatory model: causal, mechanistic, mathematical, computational, statistical, and so on.[2] If so, then the questions arise of how various explanations of both the same and different varieties are related to one another, whether we must choose among them, and how to do so. This question will be acute if the various explanatory models intend to offer a plurality of explanations of the same phenomenon, or if they explain different phenomena that are purported to "add up" in some way to the target phenomenon. This may be, for example, by standing in some sort of part-whole relation to the target phenomenon as atoms do to molecular substances, proteins to cells, individuals to populations, and so on.

These latter issues, regarding the relations that a plurality of explanations or explanatory models may stand in with respect to one another, are particularly important if we think that explanations — or some explanations, at any rate — tell us how the world is. For example, a common idea is that what exists is more or less what our best explanations tell us exists—or, to qualify, the best explanations of certain sorts. Insofar as explanations are our best guide to the world around us, we have an interest in what

[2] We shall use "model" and "explanation" interchangeably, for present purposes.

makes for a good, adequate, or otherwise successful explanation.[3] So explanations appear to have implications for what is thought to exist in the world—what philosophers call "ontological commitments."

Of course, many scientists and philosophers are unmoved by "ontological" considerations. But here, again, some things can be said. For example, insofar as we take it that explanations tell us about the world, determining just what an explanation tells us about the world (i.e., its "ontological commitments") can have implications for whether an explanation is successful at all. One example that has been recently debated is the Model-to-Mechanism Mapping (3M) constraint, according to which a model is only explanatory if the parts of the model correspond to parts of the system that it is modeling (Kaplan, 2011; Kaplan and Craver, 2011). For example, it has been argued that the Hodgkin-Huxley (HH) model of the formation of the action potential was not itself explanatory, and the explanation of the formation of the action potential was not had until the discovery of the existence and mechanism of ion channels (Craver, 2006) that correspond to the variables in the HH model.

The 3M constraint implies that an explanatory model including idealized parts that do not map directly onto the target phenomenon will always be inferior to any explanation that includes only components that map onto parts of the phenomenon; and 3M is sometimes interpreted to imply that an explanation that has more parts that map

[3] We need not suppose that this broadly realist approach to explanation is committed to any deep "metaphysical" thesis in the sense that has been debated over the last century or more. It is enough that we take explanation to be a, perhaps fallible, guide to the claims that we accept, count as true, assess to be warrantedly assertable, or whatever pro-epistemic status we prefer.

Nor should we suppose that this thin realism implies that we can read our ontological commitments directly off the quantificational structure of sentences used in explanatory texts (cf. Quine, 1948).

onto the parts of the phenomenon will always be a better explanation than those that have fewer mappings (the principle of *More Details are Better*, per Chirimuuta, 2014). But these two implications are rejected by many scientists and philosophers of science, and even by some whose accounts of explanation are purported to have these implications (e.g., Craver and Kaplan, 2020). As pointed out, the term "mechanism" in neuroscience and closely-related domains can refer to different kinds of explanations ranging from those in which the activity of a single mechanism component is explanatory, to those for which the organization or relationships between mechanism components seem to be explanatory (Konsman, 2024; Ross and Bassett, 2024). The question of whether or not the organization or relationship between the parts could be an additional criterion by which to evaluate the 3M constraint may be of relevance here because many mechanisms in cognitive science and neuroscience seem to correspond to more abstract mechanisms (Anand and Mande, 2022) — or to what some philosophers call "mechanism schemas" (Machamer et al., 2000).

The common-sense "realist" relationship between explanation and the world is hard to break but also hard to establish rigorously. Here we suggest only that the "ontological commitments" of an explanation — what it says about things in the world — bear not only on its success or failure in explanatory terms but also bear on experimental and clinical interventions. We cannot experimentally or clinically intervene on things that do not exist; and the adequacy of an explanation is sometimes assessed by the interventions that it predicts or enables. This is the sort of thought that lies behind Ian Hacking's slogan, "if you can spray them then they are real" (Hacking,

1983: 23). Hacking had in mind electrons and the use of electron "guns" in, e.g., television sets and electron microscopes.

Finally, there is a special sort of practical relevance to so-called scientific ontology that is typically neglected by philosophers but is often primary for scientists—namely, whether something exists, or whether we can find out what exists, is often relevant to decisions about the utility and funding of various research programs. For just one example that is salient in the cognitive and brain sciences, if neuroimaging techniques like functional magnetic resonance imaging (fMRI) do not contribute to explanations, then we may be wasting huge amounts of time and money on neuroimaging studies—time and money that would be better spent on other research programs. It is plausible to think that neuroimaging studies contribute to explanations only if they are evidence about the working parts of neural systems, ideally if the brain areas that "light up" in imaging studies are the working parts of brains or are closely correlated with them. Imaging studies might be useful if they were even merely heuristic for whether and where to direct other inquiries, for example, studies in systems or cellular neurobiology. But even then, many researchers would wonder if that benefit justified the costs of the studies. Such questions have in fact been raised. Indeed, it has been charged that neuroimaging—i.e., cognitive neuroscience—is wholly bankrupt because the operations of brains will be entirely explained by molecular biochemical models (Shulman, 2013) or because brains are wholistic systems (Hardcastle and Stewart 2002). This would be the case, for example, if the working parts of brains are not "brain areas" and that way of talking is just a heuristic way to refer to the gross

anatomical locations of the electrochemical processes that are the real explainers; or if explanation "drains down" to the "lowest" or "smallest" explanatory models.

Many scientists and philosophers have strongly held views about the outcomes of the above sorts of disputes. And those views likely depend on assumptions about explanation and the relationship between explanation and ontology — whether or not such assumptions are made explicit.[4] So there is good reason to think that philosophers and scientists ought to care about the nature of explanation in cognitive and brain sciences, and thus about scientific explanation in general. We shall proceed on that basis.

Varieties of Explanatory Models

As noted above, there is a plentiful number of competing proposals about the nature of explanation or explanatory models in the cognitive and brain sciences. And in the preceding discussion, we raised the possibility that there could in fact be a plurality of successful kinds of explanations or explanatory models. Indeed, w favor some varieties of explanatory pluralism but we don't intend to argue for explanatory pluralism here. For the purposes of inquiring about computational explanation in the cognitive and brain sciences, it is enough that there are multiple *candidate* accounts of scientific explanation regardless of whether they are ultimately compatible with or competing with one another.

[4] John Bickle has sometimes claimed that these issues are purely descriptive and can be resolved by accurately describing what neuroscientists do without any assumptions about what makes some explanations successful and others not, and without any appeal to ontological considerations (Bickle, 2003). For example, he points to the amount of space dedicated to molecular neuroscience posters versus cognitive neuroscience posters at meetings of Society for Neuroscience as evidence that "real" neuroscience is molecular neuroscience. It is hard to see how this can be an entirely descriptive claim.

Above we justified the project of explaining explanations in part by appealing to the potential ontological significance of explanations and explanatory models, viz., that "what exists is more or less what our best explanations tell us exists." But we immediately qualified our appeal to even this minimal realism with the observation that it may be only some kinds of explanation that have the potential to tell us what exists. So we have thus far said nothing about how different explanations could tell us about the world. Let us begin with a relatively familiar way of making the connection between explanation and the furniture of reality.

When we were offering reasons that one might want an account of explanation, we appealed to some classic examples that seem to illustrate important features of explanation. There seems to be a difference between explaining the length of the shadow in terms of the height of the flagpole and explaining the height of the flagpole in terms of the length of the shadow; and this seems to illustrate that explanatory relations are, or can be, asymmetric. There seems to be a difference between explaining Josephina's non-pregnancy and Joseph's non-pregnancy in terms of their use of contraception; and this seems to illustrate that potential explanatory factors can be distinguished according to differences in relevance. A common way to secure this explanatory asymmetry and relevance is to require explanations to appeal to a relation in the world that is itself asymmetric and differentially sensitive to manipulations or interventions. The primary candidate for such a relation is causation. We can say that the flagpole height explains the shadow length but not vice versa because the flagpole *causes* the shadow and not vice versa. And the use of contraception explains Josephina's non-pregnancy and not Joseph's if contraceptives are *causes* of Josephina's non-pregnancy and not causes of

Joseph's non-pregnancy. The idea, then, is that explanations succeed by identifying causal relations. Causal explanations, i.e., causal models, successfully explain the occurrence of effects by identifying their actual causes.

That causal explanation can account for explanatory asymmetry and relevance is often cited in favor of causal explanation as the basic sort of scientific explanation (cf. Woodward, 2004). Be that as it may, what is crucial for our purposes is that the above line of reasoning seems to say that causal explanations succeed when the world is as the explanation says, that is, when there are causal relations that correspond to those cited in a causal explanation or included in a causal model.

Now, whatever the causal relation itself might be, it is plausible that the things that stand in causal relations must exist.[5] So causal explanations tell us what exists by telling us about the relata of causal relations, where it is implied that those relata are "things" in some sense, i.e., they exist. Indeed, one widely accepted principle is that "to be is to be a cause." Jaegwon Kim calls this *Alexander's Dictum*, and others have called it the *Eleatic Principle* (Kim, 2006; Armstrong, 1978). What exists are causes.

If we take causal models of explanation accordingly, we get an easy path from explanation to existence. Put in terms of explanatory models:

1. Model M is the best causal model of phenomenon P; and, according to M, C causes E
2. Therefore, C causes E (from 1, by minimal realism)
3. Therefore, C exists (from 2, by Alexander's Dictum)

[5] Nothing has been said here about the nature of causation, but the line of reasoning is most persuasive if causation is more than mere regularity.

Once we've gone down this path of reasoning, we may draw further practical and epistemic conclusions, such as: cause C is the kind of thing on which we can experimentally or clinically intervene, cause C and model M are worthwhile objects of investigation, or M is accurate, true, or otherwise epistemically good. These further inferences begin from the existence of cause C, whatever it may be.

Of course, there are many missing details and unarticulated assumptions in the above argument outline. And we do not want to leave anyone with the mistaken impression that the path from causal explanation to ontology is uncontroversial. What is important, for present purposes, is simply that we have a good sense of how the argument would work for causal explanation, assuming it does.

This is important precisely because there are many other kinds of proposed explanation and explanatory models that are not causal models, or at least are *prima facie* not causal models. Classically we might distinguish causal explanations from teleological explanations, or proximate causes from ultimate causes (Mayr, 1961). In the second half of the twentieth century the main contenders were influentially grouped together into the classes of causal explanations and unification explanations (Salmon and Fagot-Largeault, 1989, Salmon, 1990). In the early twenty-first century it is familiar to frame questions about explanation in terms of explanatory models, where those include causal models, mechanistic models, mathematical models, physical models, computational models, statistical models, and topological models.[6] And there has also

[6] Here we follow the fashion of focusing attention on so-called explanatory models. But it is just as important—perhaps even more important in the neuroscience—to attend to other kinds of models: data models, experimental models, predictive models, clinical models, and so on.

been a great deal of attention to the ways that models — the best models, that is — do not invariably work by telling us how the world is but instead rely on strategies such as abstraction and idealization. Michael Weisberg, for example, describes three kinds of idealized models that he calls Galilean idealization, minimalist idealization, and multiple model idealization; and others have proposed even more varieties (Weisberg, 2007a).

The upshot is that, on the one hand, we cannot blithely assume that all explanations are causal, nor that figuring out what a model tells us about the world is a straight-forward matter. On the other hand, we do have a good general idea of how causal explanations work, what makes them successful, and how they are related to ontology—all of this being in principle, while recognizing that it is often difficult to determine in practice.

This brings us back to our eponymous question, namely, how do computational models in the cognitive and brain sciences explain? For we are now in a position to give a conditional answer: If computational models in the cognitive and brain sciences are causal models, then they explain by representing the causal relations between computational states, properties, events, or processes; and they succeed when there are (i.e., there exist) states, properties, events, or processes that are causally related as the model says that they are. In short, if these computational models are causal models, then computational models succeed when they get the causal relations right.[7]

[7] We use 'state' in this discussion and hereafter to include (the having of) properties and (the occurrence of) events or processes; and we shall use all of these expressions interchangeably, along with 'thing' in the most generic sense. Moreover, as far as we're concerned, a causal model or explanation can succeed or "get the causal relations right" even if it is idealized or abstract. That view is not trivial; but we do not rely on it here except for ease of exposition. The defense of that view is beyond the scope of this paper.

Taking stock of the situation, we find ourselves at a crossroads. If computational models in the cognitive and brain sciences are causal models, then in general terms we know how they do their explanatory work and what ontological implications they carry. But if computational models in the cognitive and brain sciences are not causal models, then there is more work to be done if we are to understand how they do their explanatory work and what ontological implications they carry.

Two potential sources of confusion must be mentioned before we proceed. First, just as there is a generic way of talking about computation as any mediating process whatsoever, there is also a way of talking computation that simply redescribes the phenomenon that is to be explained. We therefore need to be careful to attend to only those cases where the computational explanations are more substantive (Cao 2019).

Second, while we should be aware of extremely generic notions of computation that make our question seem too trivial to answer, we must also be alert to overly stringent notions of computation. Nico Orlandi, for example, argues that vision is not a computational process in part by arguing that processes only count as computational if they actually represent the rules that they follow (Orlandi, 2014). This idea has a long history, going back to Alan Turing's (1936, 1950) specification of his eponymous machine as composed in part by the machine table that records the instructions that are to be followed by the processing unit. But this is an extremely demanding way of thinking about computation that would rule out many familiar but simple electronic computing devices as well as, possibly, even general purpose computing machines that

run compiled programs[8] We are not aware of any model in the cognitive and brain sciences that assumes that the system to be explained itself represents computational instructions that it consults during processing. If that is what we require, then there are not even any candidates for computational explanation in the cognitive and brain sciences. But that conclusion conflicts with the practice of giving computational explanations in the brain sciences. So, it seems that we should not be quite so demanding about computation.

At this point our inquiry threatens to become bogged down in variety and ambiguity. It is well known that various notions of computation are at play, e.g., Aizawa (2010), Milkowski (2013), Piccinini (2015, 2020), Shagrir (2021), or Maley (2022). Plainly this variety could itself contribute to confusion over the nature of computational explanations.

If the question is whether there is *some* example of "computational" explanation in the cognitive and brain sciences that can be assimilated to causal explanation, then the answer is surely yes. As Piccinini and Scarantino observe, "In many quarters, especially neuroscientific ones, the term 'computation' is used, more or less, for whatever internal processes explain cognition" (Piccinini and Scarantino 2010: 244, 2011: 4; see also Cao 2019). Used so generically, it is likely that many such "computational" explanations are causal explanations. But here the term 'computes' is like the term 'mediates" and can be

[8]There is a tradition of thinking of computer programs in terms of the human-readable programs that we write; and of compiling a human-readable computer program as one of translating those same instructions into a "machine language" program that the machine will follow. But insofar as the machine program can be thought of in terms of instructions at all, it is more accurate to think of the compiled program as providing instructions for how to arrange the initial conditions of the machine for its run, rather than as a set of instructions that are followed during the running of the program.

used to refer to processes, often unknown, that are causal mediators between input and output.

Similarly, and not unrelatedly, if the question is whether there is *some* notion of computation according to which some or all causal processes in the cognitive and neural systems are computational processes, then again the answer is surely yes. That consequence would be assured by various theories of computation that imply that every process is a computational process, i.e., pancomputation. But it would also be implied by a variety of theories of computation according to which whether something is a computation is a matter of interpretation or description, an idea to which we shall return. These questions are important, but they do not tell us what we want to know about computational explanation in the cognitive and brain sciences.

Fortunately for us, we do not have to settle on one example of computational explanation nor one theory of computation. What we want to know is not, in the first case, whether one or more of the accounts is correct. Rather, we can turn directly to some of the proposed accounts of computational explanation in the cognitive and brain sciences and consider whether they purport to be causal accounts or to make claims about ontology. That is the task of the next section.

Are computational models in fact causal models?

Are computational models in the cognitive and brain sciences in fact causal models? One answer is, yes. According to Gualtiero Piccinini (Piccinini, 2015, 2007), computational models are abstract mechanical models, viz., mechanism sketches or schema. He specifies his "mechanistic account of computation" as so:

> A physical computing system is a mechanism whose teleological function is performing a physical computation. A physical computation is the manipulation (by a functional mechanism) of a medium-independent vehicle according to a rule. A medium-independent vehicle is a physical variable defined solely in terms of its degrees of freedom (e.g., whether its value is 1 or 0 during a given time interval), as opposed to its specific physical composition (e.g., whether it's a voltage and what voltage values correspond to 1 or 0 during a given time interval). A rule is a mapping from inputs and/or internal states to internal states and/or outputs. (Piccinini, 2015)

Piccinini is explicit that his account is meant to ensure that computational explanation is just a special case of causal explanation — specifically, of mechanistic explanation (2015, 2020). A similarly "mechanistic" account of neural computation has been advanced by Milkowski (2013).

Whether Piccinini's "mechanistic account of computation" should be favored is not our current question.[9] What matters for our present purposes is simply that *if* Piccinini is correct then there is no special problem about computational explanation in the brain sciences — because computational explanation is just a special case of mechanistic explanation, which is itself just a variety of causal explanation. We argued above that, at least in outline, philosophers and scientists understand what sorts of scientific practices should be used to explore such mechanisms, and we understand what sorts of scientific products will count as successes. So there is no special problem about computational explanation in the brain sciences if Piccinini is correct.

[9] For doubts, see Shagrir (2021) and Maley (2023).

A different answer is also affirmative but it works from concrete examples rather than by employing an overarching mechanistic assumption. As Piccinini and Scarantino say: "Many neuroscientists have started using the term 'computation' for the processing of neuronal spike trains (i.e. sequences of spikes produced by neurons in real time). The processing of neuronal spike trains by neural systems is often called 'neural computation'" (Piccinini and Scarantino, 2011). While there is reason to doubt that the processing of neuronal spike trains is best understood as a variety of computation (Cao 2019), there is no doubt that such "neural computation" is a causal process. So, as above, we may comfortably conclude that this variety of computational explanation in the brain sciences can be assimilated to causal explanation.

A third approach, advanced by Mazviita Chirimuuta (2014, 2018), proposes that computational explanation in brain sciences is distinct from the mechanistic account of explanation. According to Chirimuuta, models in computational neuroscience often produce a non-mechanistic type of explanation which she dubs *efficient coding explanation*. Efficient-coding explanations rely on the idea that neural systems are structured to represent information in a way that minimizes redundancy and maximizes the use of available resources, such as energy or neural resources (Doi et al., 2012; Evan C. Smith et al., 2006). This idea derives from information theory, which suggests that neural systems might encode sensory information in an efficient manner to optimize processing and transmission. Efficient coding explanations, Chirimuuta argues, are best viewed as a variety of minimal model that differs from Weisberg's causal minimal models (Weisberg, 2007b) which she calls 'A-minimal models' (Chirimuuta, 2014, p. 134). Whereas A-minimal models operate in the mechanistic mode, seeking to identify

some specific causal mechanisms underlying cognitive and neural processes (or "difference makers", per Levy and Bechtel, 2013), efficient-coding models operate in the computational mode, aiming to describe how information is encoded and processed in the neural system by focusing on abstract principles of optimization and information processing. Building upon an influential efficient-coding model that abstracts from the biophysical mechanisms underlying its neural implementation (viz., the Gabor model of V1 receptive field), Chirimuuta contends that despite being basically descriptive, the model offers a computational explanation for why neurons exhibit behaviors described by the model. The Gabor model of V1 receptive fields abstracts away from biophysical details not because it is incomplete or because the details are unknown, but instead it abstracts in order to characterize the information-processing capabilities of a neuron or neuronal population. In other words, efficient-coding models do not try to satisfy the 3M or the MDB ("more details are better") constraints.

If 3M is constitutive of mechanistic explanation, then efficient-coding explanations are not mechanistic. Yet, they nevertheless offer genuine explanations, as Chirimuuta argues by showing that efficient-coding computational explanations can meet the same explanatory demands endorsed by the advocates of mechanistic explanation — such as James Woodward's counterfactual test for causal explanation. In particular, efficient-coding models can answer "w-questions" or "what-if-things-had-been-different questions"(Woodward, 2004). Indeed, the facts that efficient-coding explanations address the processing of information in the neural system and that they are dependent on the evolutionary or developmental environments is precisely what allows them to offer specific answers to questions about what would occur in counterfactual

scenarios that modulate the behavior of the system. On Chirimuuta's account then, at least some computational models in brain sciences are not mechanistic while still being explanatory by the same criteria as causal models[10].

Chirimuuta's approach reflects on the practice of neuroscience, noting that it displays a wider diversity of methods of inquiry than is usually considered by philosophers. If she is on the right path, the upshot is that there are several types of explanation at play in the brain sciences, depending on the type of abstraction "level" (rather than scale or constitutive "levels") at which the explanation is directed. Some of these explanations are not causal or mechanical. But if Chirimuuta is right then these explanations only require us to slightly relax our explanatory criteria. She argues that efficient-coding explanations, for example, satisfy broadly Woodwardian criteria on explanation, such as that they can answer counterfactual questions about why things are the way they are and how they could be different. Suppose that the core idea behind "Alexander's Dictum" that what exists are causes is the idea the things that exist are the things about which we can reason counterfactually, e.g., to answer such "w-questions" or "what-if-things-had-been-different questions"(Woodward, 2004). This suggests that we might generalize the principle to something like, "what exists are those things on which what-if-things-had-been-different questions depend." If so, then we can assimilate Chirimuuta's efficient-coding models into our general understanding of the methodological and ontological significance of causal explanations, now understood as dependence explanations more broadly (cf. Bechtel and Shagrir 2015).

[10] Chirimuuta later (2018) introduces the idea that "if one is willing to extend the notion of a mechanism to include the whole apparatus of natural selection and ontogenesis, one might propose that computational neuroscience is still just in the business of discovering mechanisms" (2018: 853). But that stretches the notion of a mechanism rather thin.

A different approach is advanced by Lauren Ross. Like Chirimuuta, Ross is critical of the broadness of the hegemony of mechanistic explanation in neuroscience and of the 3M criterion (Ross, 2015; Ross and Bassett, 2024). She proposes that computational explanations in the brain sciences are better construed as explanations in terms of causal circuits or pathways rather than as mechanisms, in a special sense of Craver (Craver, 2007) or even (Glennan and Illari, 2018). Mechanisms in this view "emphasize the "biophysical" and "physical" causes that realize neural and brain systems," which can be "called the hardware or wetware" and that is distinguished "from higher-scale structures, computations" (Ross and Bassett, 2024, 84). While this does not exclude a 3M approach as such, it is certainly not firmly committed to it. Instead, Ross seems more favorable to a Chirimuuta-style approach of computational explanations that "pertains not just to single neurons but also to neural networks" and thus "indicates the relevance of this explanatory approach to both cellular- and systems-level neuroscience" (Ross, 2015). If Ross is right and computational models in the brain sciences are actually circuit or pathway models, then once more we have a view that assimilates computational explanation to causal explanation, albeit not mechanistic explanation per se. In that case we at least know how to think about the ontological commitments and practical implications of circuits and pathways.

Up to this point, we seem to be suggesting that the leading interpretations of computational explanation in the cognitive and brain sciences can all be assimilated into the broadly causal or dependency framework associated with Woodward (2004) and therefore are only slight challenges to the dominant mechanistic approach to explanation in neuroscience. You might even wonder why we bothered to argue that there are

important implications for understanding different accounts of explanation if they can all be unified. But matters are not so simple.

Chirimuuta (2014) and Ross (2023) each suggest that some of the explanations they consider can not be assimilated into the broadly causal picture by itself. Ross focuses on explanations that are formulated in terms of constraints on systems, which is a familiar enough idea. She argues that some constraints are physical (like a river bed that constrains the flow of water), some have to do with laws of nature (like the constraints on body shapes and sizes for aquatic animals), and some are mathematical (like constraints on possible unbroken and non-redundant paths that cross the bridges of a city). Causal and nomological (i.e., lawful) constraints are not (always) mechanisms, but it is readily shown that they meet the same kinds of criteria on explanation that mechanists endorse. But mathematical explanations are of a different sort, and some computational explanations in neurosciences seem to involve mathematical constraints, specifically, constraints on the efficient transmission or processing of information. These will not count as full constraint explanations if they do not entirely rule out possible forms, showing that some are impossible. Instead, like the explanation of the shape of the cells of a honeycomb (Lyon, 2012), such explanations propose to explain a phenomenon by showing that it is or approximates an optimal form. Some of Chirimuuta's efficient-coding explanations seem to be of this sort.

Chirimuuta (2018) decouples the counterfactualist from the causal interventionist components of Woodward's account of explanation (Woodward, 2004) and combines the former with Lange's (2016) account of "distinctively mathematical explanation" to describe a subset of efficient-coding explanations in computational neuroscience that she

considers neither mechanistic nor causal. As she puts it in her concluding section: "The clearest cases of non-causal explanation in neuroscience are efficient coding explanations that refer to information theoretic trade-offs in order to show why it is that neural systems should employ particular computational solutions, such as hybrid computation, or Gabor filtering" (2018: 875). These kinds of computational models offer mathematical explanations of the efficiency and utility of features of neural systems because they focus on characteristics of the global dynamics of the system rather than looking for any actual causal interactions between parts of a real neural network. She suggests that this entails a perspectivist understanding of explanation in brain science which shouldn't be a surprise because neurons can be considered as complex biological mechanisms (or parts of systems), as well being considered as computational systems (or parts of systems). While acknowledging the limitations of efficient-coding explanations (particularly their reliance on simplifying assumptions and idealized models of neural processing), Chirimuuta suggests that efficient-coding complements other approaches that consider the complexity and heterogeneity of actual neural circuits.

Explanations in terms of mathematical constraints or optimality considerations do not readily fit into the picture of explanations that answer w-questions as telling us about how the world is or inviting causal intervention on the entities that they model because they frequently idealize and sometimes in dramatic ways. Sometimes their efficiency or optimality can be explained in terms of conserved resources, as perhaps is the case with the hexagonal cells of a honeycomb that require less wax to construct than would be required by any other tessellation of the same area (Lyon, 2012). But in other

cases the informational efficiency is less clear. At the very least, we can't assume that neural computational happens the way that it does because it approximates optimal energenti efficiency.

While debates persist regarding whether computational models constitute causal explanations, various perspectives converge on the idea that they are explanatory. Piccinini's mechanistic account suggests that computational explanation is a subset of causal explanation, aligning with the broader framework of mechanistic understanding in neuroscience. Chirimuuta's efficient-coding explanation, offer a non-mechanistic perspective, emphasizing optimization principles over mechanistic details. Similarly, Ross proposes a view that focuses on causal circuits or pathways rather than strict mechanistic accounts, further broadening the scope of computational explanation. These diverse interpretations collectively underline the richness and complexity of explanatory frameworks in neuroscience, challenging traditional mechanistic paradigms and urging a more nuanced understanding of causality and explanation.

It appears that computational modeling has emerged as a distinctive and influential contribution to the field of neuroscience, reshaping our understanding of explanatory norms.

Conclusion: Are computational models actually causal models, or not?

The significance of determining whether computational models are causal models for brain scientists cannot be overstated. Understanding the nature of computational explanation informs not only theoretical frameworks but also practical implications for research methodologies and funding priorities. If computational models are indeed

causal models, scientists can leverage established methodologies for exploring causal relationships to advance understanding in cognitive and brain sciences. Conversely, if computational models offer a different mode of explanation, such as efficient coding explanations, then that recognition necessitates a reevaluation of research practices and the development of new methodological approaches to capture these explanations. The most obvious implication is that computational explanations are not just more abstract or less detailed mechanistic explanations, and the question of the relationship between computational and mechanistic explanations is thrown wide open.

The long-standing priority given to the search for causal mechanisms is linked to certain types of tools (both methodological and technical) that have been crucial to the development of neuroscience as we know it today. But the increasing importance of theoretical and computational approaches in neuroscience means that we need to rethink the articulation of modes of investigation in neuroscience and the epistemic standards that go with them. As such the question of whether computational models are causal models or not has profound implications for the advancement of knowledge in cognitive and brain sciences. Answering this question enables scientists to refine their theoretical frameworks, develop robust research methodologies, and ultimately deepen our understanding of the complexities of the human brain.

Acknowledgments

REMOVED FOR ANONYMOUS REVIEW

Ethics, Animal Care, and Data Sharing

This is a theoretical study. No experiments were conducted and no data were produced.

References

Altimus, C.M., Marlin, B.J., Charalambakis, N.E., Colón-Rodriquez, A., Glover, E.J., Izbicki, P., Johnson, A., Lourenco, M.V., Makinson, R.A., McQuail, J., Obeso, I., Padilla-Coreano, N., Wells, M.F., for Training Advisory Committee, 2020. The Next 50 Years of Neuroscience. J Neurosci 40, 101–106. https://doi.org/10.1523/JNEUROSCI.0744-19.2019

Armstrong, D.M., 1978. A Theory of Universals. Universals and Scientific Realism Volume Ii. Cambridge University Press.

Bechtel, W., 2009. Looking down, around, and up: Mechanistic explanation in psychology. Philosophical Psychology 22, 543–564.

Bechtel, W., 2008. Mental mechanisms: philosophical perspectives on cognitive neuroscience. Psychology Press, New York.

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

Bickle, J., 2003. Philosophy and Neuroscience: A Ruthlessly Reductive Account. Kluwer Academic Publishers.

Boone, W., Piccinini, G., 2016. The cognitive neuroscience revolution. Synthese 193, 1509–1534.

Bromberger, S., 1966. Why-Questions, in: Colodny, R.G. (Ed.), Mind and Cosmos -- Essays in Contemporary Science and Philosophy. University of Pittsburgh Press, pp. 86--111.

Cao, R., 2019. Computational explanations and neural coding.in Routledge Handbook of theComputationMind,(ed. Sprevak& Colombo.

Casadevall, A., Fang, F.C., 2008. Descriptive science. Infect Immun 76, 3835–3836. https://doi.org/10.1128/IAI.00743-08

Chirimuuta, M., 2018. Explanation in Computational Neuroscience: Causal and Non-causal. British Journal for the Philosophy of Science 69, 849–880.

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

Claxton, K., Cohen, J.T., Neumann, P.J., 2005. When is evidence sufficient? Health Aff (Millwood) 24, 93–101. https://doi.org/10.1377/hlthaff.24.1.93

Cooper, R.P., Shallice, T., 2010. Cognitive neuroscience: the troubled marriage of cognitive science and neuroscience. Topics in cognitive science 2. https://doi.org/10.1111/j.1756-8765.2010.01090.x

Craver, C., 2007. Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience. Oxford University Press, Oxford.

Craver, C.F., 2006. When mechanistic models explain. Synthese 153, 355–376. https://doi.org/10.1007/s11229-006-9097-x

Craver, C.F., Kaplan, D.M., 2020. Are More Details Better? On the Norms of Completeness for Mechanistic Explanations. British Journal for the Philosophy of Science 71, 287–319.

Doi, E., Doi, E., Gauthier, J., Gauthier, J.L., Field, G.D., Field, G.D., Shlens, J., Jonathon Shlens, Shlens, J., Jonathon Shlens, Sher, A., Sher, A., Greschner, M., Greschner, M., Machado, T.A., Machado, T.A., Jepson, L.H., Jepson, L.H., Mathieson, K., Mathieson, K., Gunning, D.E., Gunning, D.E., Litke, A. M., Litke, Alan M., Litke, Alan M., Litke, A., Paninski, L., Liam Paninski, Chichilnisky, E.J., Chichilnisky, E.J., Simoncelli, E.P., Simoncelli, E.P., 2012. Efficient Coding of Spatial Information in the Primate Retina. The Journal of Neuroscience 32, 16256–16264. https://doi.org/10.1523/jneurosci.4036-12.2012

Evan C. Smith, McClelland, J.L., Michael S. Lewicki, Lewicki, M.S., Smith, E.C., 2006. Efficient auditory coding. Nature 439, 978–982. https://doi.org/10.1038/nature04485

Glennan, S., Illari, P., 2018. Introduction: mechanisms and mechanical philosophies, in: Glennan, S., Illari, P. (Eds.), The Routledge Handbook of Mechanisisms and Mechanical Philosophy. Taylor & Francis Group, Oxford, UK, p. 476.

Hacking, I., 1983. Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge University Press.

Hanna, J.F., 1969. Explanation, prediction, description, and information theory. Synthese 20, 308–334.

Kaplan, D.M., 2011. Explanation and description in computational neuroscience. Synthese 183, 339–373.

Kaplan, D.M., Craver, C.F., 2011. The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. Philos. Sci. 78, 601–627. https://doi.org/10.1086/661755

Kim, J., 2006. Emergence: Core ideas and issues. Synthese 151, 547–559.

Konsman, J.P., 2024. Expanding the notion of mechanism to further understanding of biopsychosocial disorders? Depression and medically-unexplained pain as cases in point. Studies in History and Philosophy of Science Part A 103, 123–136.

Kosslyn, S.M., Shin, L.M., 1992. The status of cognitive neuroscience. Curr Opin Neurobiol 2, 146–149. https://doi.org/10.1016/0959-4388(92)90002-3

Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., Poeppel, D., 2017. Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron 93, 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Kriegeskorte, N., Douglas, P.K., 2018. Cognitive computational neuroscience. Nat Neurosci 21, 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Lange, M., 2016. Because Without Cause: Non-Causal Explanations in Science and Mathematics. Oxford University Press, Oxford. UK.

Lange, M., 2013. What Makes a Scientific Explanation Distinctively Mathematical? British Journal for the Philosophy of Science 64, 485–511.

Levy, A., Bechtel, W., 2013. Abstraction and the Organization of Mechanisms. Philosophy of Science 80, 241–261. https://doi.org/10.1086/670300

Lyon, A., 2012. Mathematical Explanations of Empirical Facts, and Mathematical Realism. Australasian Journal of Philosophy 90, 559–578. https://doi.org/10.1080/00048402.2011.596216

Machamer, P., Darden, L., Craver, C.F., 2000. Thinking about mechanisms. Philos. Sci. 67, 1–25. https://doi.org/10.1086/392759

MacQueen, G., 2013. What does it mean to have enough evidence? J Psychiatry Neurosci 38, 3–5. https://doi.org/10.1503/jpn.120240

Mayr, E., 1961. Cause and effect in biology. Science 134, 1501–6. https://doi.org/10.1126/science.134.3489.1501

Milkowski, M., 2013. A Mechanistic Account of Computational Explanation in Cognitive Science, in: Knauff, M., Pauen, M., Sebanz, N., Wachsmuth, I. (Eds.), Cooperative Minds: Social Interaction and Group Dynamics. Proceedings of the 35th Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 3050–3055.

Orlandi, N., 2014. The Innocent Eye: Why Vision is Not a Cognitive Process. Oup Usa.

Pereira, A., 2007. What the cognitive neurosciences mean to me. Mens Sana Monogr 5, 158–168. https://doi.org/10.4103/0973-1229.32160

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

Piccinini, G., 2007. Computing mechanisms. Philosophy of Science 74, 501–526.

Piccinini, G., Scarantino, A., 2011. Information processing, computation, and cognition. Journal of Biological Physics 37, 1–38.

Potochnik, A., Sanches de Oliveira, G., 2020. Patterns in Cognitive Phenomena and Pluralism of Explanatory Styles. Top Cogn Sci 12, 1306–1320. https://doi.org/10.1111/tops.12481

Quine, W.V.O., 1948. On What There Is. Review of Metaphysics 2, 21–38.

Reese, H.W., 1999. Explanantion is not description. Behavioral Development Bulletin 8, 3–7. https://doi.org/doi.org/10.1037/h0100524

Ross, L.N., 2023. The explanatory nature of constraints: Law-based, mathematical, and causal. Synthese 202, 56. https://doi.org/10.1007/s11229-023-04281-5

Ross, L.N., 2015. Dynamical Models and Explanation in Neuroscience. Philosophy of Science 82, 32–54.

Ross, L.N., Bassett, D.S., 2024. Causation in neuroscience: keeping mechanism meaningful. Nat Rev Neurosci 25, 81–90. https://doi.org/10.1038/s41583-023-00778-7

Salmon, W.C., 1971. Statistical explanation & statistical relevance. University of Pittsburgh Press.

Salmon, W.C., Fagot-Largeault, A., 1989. Four Decades of Scientific Explanation. History and Philosophy of the Life Sciences 16, 355.

Shulman, R.G., 2013. Brain imaging. Oxford University Press.

Weisberg, M., 2007a. Three Kinds of Idealization. Journal of Philosophy 104, 639–659.

Weisberg, M., 2007b. Who is a Modeler? British Journal for the Philosophy of Science 58, 207–233.

Woodward, J., 2004. Making Things Happen. https://doi.org/10.1093/0195155270.001.0001