Revista Argentina de Ciencias del Comportamiento

ISSN 1852-4206

Abril, 2017, Vol. 9, N°1, 9-18

revistas.unc.edu.ar/index.php/racc

Mechanism, Explanatory Pluralism and Efficient Coding Explanation in Neuroscience

Barberis, Sergio Daniel*,[a]

Original Article

Abstract

There is a growing debate within the Philosophy community concerning the unity and diversity of explanation in neuroscience. The new Mechanist philosophy claims that neuroscience exhibits a mosaic unity in which models from multiple scientific fields contribute to the collective mechanistic explanation of an explanandum phenomenon φ by setting causal constraints on the space of possible mechanisms for φ. Non-mechanist philosophers acknowledge the relevance (even the centrality) of mechanistic research, but they want to emphasize the plurality and diversity of explanatory research programs in neuroscience. In this paper I argue, first, that the kind of explanatory pluralism many non-mechanist philosophers endorse —which I call 'causally restricted pluralism'— is not a genuine alternative to Mechanism. Then, I present a liberalized interpretation of explanatory pluralism, one according to which there are models in neuroscience that contribute to the collective explanation of some phenomenon φ but that are not intended to set causal constraints on the space of possible mechanisms for φ. Finally, I review an explanatory research program in neuroscience, namely, efficient coding explanation, which is better accounted for by the liberalized interpretation of pluralism.

Key Words:
Mechanism, Unity of Science, Explanatory Pluralism, Causality, Efficient Coding Explanation,

Resumen

Mecanicismo, Pluralismo Explicativo y Explicación de Codificación Eficiente en Neurociencia. Hay un debate creciente en la comunidad filosófica acerca de la unidad y la diversidad de la explicación en neurociencia. La nueva filosofía mecanicista sostiene que la neurociencia exhibe una unidad de mosaico, según la cual los modelos provenientes de múltiples campos científicos contribuyen a la explicación mecanicista colectiva de un fenómeno explanandum φ mediante el establecimiento de restricciones causales sobre el espacio de mecanismos posibles para φ. Los filósofos no mecanicistas admiten la relevancia (incluso la centralidad) de la investigación mecanicista, pero enfatizan la pluralidad y la diversidad de los programas de explicación en neurociencia. En este artículo argumento, en primer lugar, que el tipo de pluralismo explicativo que muchos filósofos no mecanicistas defienden —lo que llamo el "pluralismo causalmente restringido"— no es una alternativa genuina al mecanicismo. Luego, presento una interpretación liberalizada del pluralismo explicativo, según la cual existen modelos en neurociencia que contribuyen a la explicación colectiva de un fenómeno φ pero que no pretenden establecer restricciones causales sobre el espacio de mecanismos posibles para φ. Finalmente, reseño un programa de explicación en neurociencia, a saber, la explicación de codificación eficiente, que se entiende de manera más adecuada mediante la interpretación liberalizada del pluralismo.

Palabras clave:
Mecanicismo, Unidad de la Ciencia, Pluralismo Explicativo, Causalidad, Explicación de Codificación Eficiente.

Tabla de Contenido

Received on June 8th, 2016; Review received on October 17th, 2016; Approved on October 19th, 2016.
This article was edited by: Ricardo Pautassi, Gabriela Rivarola, Débora Jeanette Mola, Daniela Alonso.

1. Introduction

There is a growing debate within the Philosophy community concerning the unity and diversity of explanation in neuroscience. The new Mechanist philosophy claims that neuroscience exhibits a mosaic unity, one in which models from diverse scientific fields contribute to the collective explanation of an explanandum phenomenon φ, by setting causal constraints on the space of possible mechanisms (hereafter: SPM) for φ. For example, some mechanist philosophers claim that 'cognitive science', as traditionally conceived (cf. Marr, 1982), is on its way out and is being replaced by 'cognitive neuroscience', an interdisciplinary scientific field that aims to build multilevel mechanistic explanations for cognitive phenomena (Boone & Piccinini, 2016a; 2016b) see also Piccinini & Craver, 2011).

[a] Universidad de Buenos Aires, Facultad de Filosofía y Letras, Instituto de Filosofía: "Alejandro Korn". Buenos Aires. Argentina.
*Enviar correspondencia a: Barberis, S. D. E-mail: sbarberis@filo.uba.ar

Citar este artículo como: Barberis, S. D. (2017). Mechanism, Explanatory Pluralism and Efficient Coding Explanation in Neuroscience. *Revista Argentina de Ciencias del Comportamiento, 9*(1),9-18

Non-mechanist philosophers acknowledge the relevance (even the centrality) of mechanistic research programs, but they want to emphasize the plurality and diversity of explanatory research programs in neuroscience (Chirimuuta, 2014; Weiskopf, 2011). They denounce that "we are in the midst of a mania for mechanisms", a mania that may lead to a kind of mechanism imperialism that "neglects the possibility that a system's behavior can be explained from many distinct epistemic perspectives, each of which is illuminating" (Weiskopf, 2011, p. 334).

In this paper, I present the ontic interpretation of Mechanism (Section 2) and argue that the ontic interpretation of Mechanism is compatible with a very appealing kind of explanatory pluralism —namely, causally restricted pluralism— that some non-mechanist philosophers have endorsed (Chirimuuta, 2014; Weiskopf, 2011). Then, I introduce a liberalized interpretation of explanatory pluralism, one according to which there are models in neuroscience that contribute to the collective explanation of some phenomenon φ but that are not intended to set any causal constraints on the SPM for φ (Section 3). By reviewing an explanatory research program in computational neuroscience, namely, efficient coding explanation, I argue that efficient coding explanation is better accounted for by the liberalized interpretation of explanatory pluralism (Section 4).

2. Mechanism

In this section, I characterize the mechanistic perspective on explanation in neuroscience. What is Mechanism about? Some philosophers think that it is primarily a thesis concerning the *vehicles* of explanation (Bechtel & Abrahamsen, 2005; Wright & Bechtel, 2007). Mechanists about the vehicles of explanation claim that the primary mode of presentation of explanations in neuroscience consists of models of the mechanism taken to be responsible for the explanandum phenomenon. A model of a mechanism aims to represent "its relevant component parts and operations, the organization of the parts and operations into a system, and the means by which operations are orchestrated so as to produce the phenomenon" (Bechtel & Abrahamsen 2005, p. 425). This thesis is intended to be an alternative to the covering-law 'model' of explanation, according to which explanation does not involve presenting a model of a mechanism but a logical inference from laws to explanandum statements.

Other philosophers think that Mechanism is not about the vehicles of explanation, that is, about the representational format that scientific explanations take in neuroscience (Craver, 2015). Mechanistic explanations may be conveyed by mechanism schemas (Machamer, Darden, & Craver, 2000), prototype vectors (Churchland, 1989), or abstract mathematical equations (Levy, 2013), just to name a few representational vehicles. Mechanism is, instead, a thesis about the truth conditions of successful explanations in neuroscience (Krickel, forthcoming). In order to make the truth conditions of explanation explicit, "one must look beyond representational structures to the ontic structures in the world" (Craver 2015, p. 28).

According to Craver (2007, p. 26), Wesley Salmon's "most penetrating insight" was to notice, first, an ambiguity in the term 'explanation'. In a sense, explanations are explanatory texts –scientific representations– that function as vehicles for conveying information about an explanandum phenomenon. As explanatory texts, explanations may be more or less speculative, empirically adequate or precise. Explanatory texts are the kind of entities that may be true or false. As epistemic products, they are "complex representations operated upon to generate knowledge and facilitate understanding" (Wright 2012, p. 376). In another sense, however, explanations are worldly structures, objective features of the world, causal-mechanical patterns and regularities into which events fit. Considered as ontic structures, explanations are not the kind of entities that may be true or false. They just are. Secondly, Salmon's approach offers insight into the idea that objective explanations (mechanisms) are the truth-makers of explanatory texts (mechanistic models). A model of a mechanism is explanatory of a phenomenon φ *in virtue* of representing some aspects of the actual mechanism that is responsible for φ. Successful models of a mechanism are 'explanatory' not because of their form but because of their true content, i.e. because they succeed in describing real aspects of the target mechanism in the world. This is what the ontic view of mechanistic explanation relies on (Craver, 2007; Glennan, 2002, 2010; Piccinini, 2007; Thagard, 2003).

Against the ontic view, Bechtel and Abrahamsen (2005, p. 424) remark that Salmon's "important insight" is that mechanisms are real systems *in* nature, and hence "one does not have to face questions comparable to those faced by nomological accounts of explanation about the ontological status of laws of nature." From an ontological point of view, mechanisms are less puzzling than laws of nature. But it is misleading to interpret this insight as implying

that the mechanism in nature directly performs the explanatory work. Bechtel (2008, p. 18) makes this point explicit when he affirms that "explanation is fundamentally an epistemic activity performed by scientists." In the same vein, Wright (2012, p. 375) affirms that "the default sense of the infinitive *to explain* is a communicative one, pertaining to the transmission of understanding" and that "explaining some phenomenon φ involves operating on internal and/or external representations of φ to understand the how or the why." From the epistemic view, to say that scientific explanations are mechanisms, or that mechanisms explain is merely to speak using metaphors of personification. Mechanisms do no explain themselves (Bechtel, 2008, p. 18). Sentences like: *Mechanistic explanations involve mechanisms* are explanatory claims in which the *representation*- and *model*-talk has been omitted from the construction. Some advocates of the epistemic conception of Mechanism emphasize that the justification of the ontic conception rests mostly on linguistic issues (Wright, 2012). Remember that Salmon's (1984) main argument in favor of the ontic conception starts off asserting the lexical ambiguity of explanation. Wright (2012) replies that explanation, and other cognate terms, are straightforwardly unambiguous. Mechanisms 'explain' only in a metaphorical or elliptical way. I think that this reason against the ontic conception is irrelevant. The ontic view of explanation does not depend on any particular analysis of the ordinary uses of explanation in natural languages. The main tenet of the ontic view is that a model of a mechanism explains a phenomenon φ to the extent, and only to the extent, that there is a real mechanism that is the truthmaker of the model and that mechanism produces φ.

I believe that the ontic view of explanation is essential to ground the normative mechanistic distinction between how-possible, how-plausible and how-actually models of mechanisms (Craver, 2006; 2007). How possibly models are not purely phenomenal models of φ (i.e. models that merely re-describe φ), but loosely constrained conjectures about the causal features of the actual mechanism that produces φ. How-possibly models may exhibit some kind of dynamical organization of parts and activities, but the modeler does not know whether those components are real or whether they are organized as the model describes. How-actually models, on the other hand, describe all and only the real parts, activities and organizational features of the mechanism that actually produce φ. In between how-possible and how-actually models there are models that vary in their degree of mechanistic plausibility. Of course, mechanists accept that every useful model introduces some distortion factors within the representation of the target system, such as idealizations, abstractions, fictions, approximations, and so on. Thus, no model can be a how-actually model of a mechanism *strictu sensu.* However, the distinction between how-possible and how-actually models provides a normative ideal in light of which explanatory *progress* in neuroscience can be identified. There is a reasonable sense in which Descartes's model of nerve action or Gall's model of brain organs were how-possibly models that failed to become explanatory because there were found to be false, that is, because the actual mechanism that would have made the explanatory claims of those models true did not exist. The ontic conception of explanation is required to make this 'reasonable sense' explicit. Models may progress in the how-possibly/how-actually *continuum* only if they represent causal components of an actual mechanism in the world.

The mechanistic conception of integration in neuroscience also demands the ontic view. Craver (2007, p. 231) claims that "the unity of neuroscience is achieved as different fields integrate their research by adding constraints on multilevel mechanistic explanations" (see also Boone & Piccinini 2016a, 2016b; Piccinini & Craver, 2011). Mechanistic collective explanations proceed through the accumulation of causal constraints from different scientific fields on the SPM for a given phenomenon φ. The SPM for φ contains all the mechanisms that could possibly explain φ (Craver, 2007). Single how-possibly models are represented by points in this space; classes of similar mechanisms are regions. A constraint is a piece of information that shapes the boundaries of the SPM or changes the probability distribution over that space (i.e. the probability that some region of the space describes the actual mechanism). The dimensionality of the SPM is fully determined by the entities, activities and organizational properties that compose the how-possibly mechanisms at all relevant levels of mechanism. The collective mechanistic explanation for the phenomenon φ is realized by the piecemeal accumulation of causal constraints on a common mechanism schema. If the target mechanism did not exist, then the whole explanatory enterprise would be wrong-headed and the diversity of scientific fields involved in that research program would not be successfully integrated.

3. Explanatory Pluralism

In this section, I argue that the ontic interpretation of Mechanism is compatible with an interpretation of explanatory pluralism —namely, causally restricted pluralism— that some non-mechanist philosophers have endorsed (Chirimuuta 2014; Weiskopf, 2011). I have already mentioned that Weiskopf (2011, p. 334) encourages a view of explanation in neuroscience in which "a system's behavior can be explained from many distinct epistemic perspectives, each of which is illuminating." According to Weiskopf's (*ibid*) version of explanatory pluralism, "[v]iewed from one perspective, the brain might be a hierarchical collection of neural mechanisms; viewed from another, it might instantiate a set of cognitive models that classify the system in ways that cut across mechanistic boundaries." Similarly, Chirimuuta (2014, p. 148) recommends a kind of explanatory pluralism "whereby the same system in neuroscience can be represented and modeled in a variety of ways". She claims that "these different perspectives on a system need not be in competition and may well be complementary" (Chirimuuta 2014, p. 148).

The main motivation behind explanatory pluralism is that neuroscience deals with a system (the brain) that is extremely complex, by almost any standard (Dale, Dietrich, & Chemero, 2009). Modelers confronted with the task of theoretically representing the structure and internal dynamics of complex systems will usually adopt an *idealization approach* to those systems (Levins, 1966; Weisberg, 2006, 2007). The idealization approach is a reasonable alternative to a more 'brute-force' approach. The latter aims to build into the model as much of the target system's complexity as possible, that is, they intend to build a model that is a "faithful, one-to-one reflection of this complexity" (Levins, 1966, p. 421). In the idealization approach, in contrast, the modeler accepts from the outset that some aspects of the explanandum phenomenon will not be incorporated into the model. In Chirimuuta's (2014, p. 149) terms, when facing the challenges of complexity, "the standard scientific response is to simplify the problem space: restrict attention to a limited range of causally significant components and forget about trying to model all of them." A natural consequence of the idealization approach is the proliferation of modeling perspectives about the target complex system, since each perspective may be useful to highlight different aspects of the complex system (Weiskopf, 2011). Explanatory pluralism is the attempt to normatively ground the proliferation of modeling perspectives that occurs naturally within the idealization approach to complex systems, such as the brain.

Weiskopf (2011) and Chirimuuta (2014) put a lot of effort into saving explanatory pluralism from the perils of *explanatory anarchism*, a hypothetical philosophy according to which 'anything goes' (Feyerabend, 1975) in neuroscientific explanation. Abney et al. (2014, p. 3) remark that "explanatory pluralism does not imply the anarchistic idea that 'anything goes': often, more than one approach is needed, but not all approaches are equally motivated, and many are even not warranted." For these authors, explanatory pluralism must be restricted somehow in order to avoid explanatory anarchism. Weiskopf (2011, p. 336, my emphasis) argues that cognitive models in cognitive psychology gain explanatory traction by picking out *real* "*strands* in the complex *causal web* that winds through the brain." Chirimuuta (2014, p. 128) argues that interpretative models in computational neuroscience "typically abstract away from many biophysical details of the neural system, in order to highlight dominant *causal influences* or universal behavior." From these quotations, one may infer that Weiskopf (2011) and Chirimuuta (2014) think that models are explanatory to the extent that they convey information about the causal factors that produce or maintain the explanandum phenomenon, whether that information is mechanistic or not. Chirimuuta (forthcoming) explicitly accepts the existence and legitimacy of non-causal explanations in neuroscience. She argues that efficient coding explanation in computational neuroscience is not causal. The motivation behind this revision is that efficient coding explanation is seen now as a kind of distinctively mathematical explanation. In this paper, I focus on Chirimuuta's (2014) arguments because I think that they express a version of explanatory pluralism that many philosophers may find appealing. In this sense, Weiskopf (2011, p. 335) asserts that cognitive models describe "real parts" of the complex causal system that produces the phenomenon, although the parts represented in the model are not "mechanistic" components of the system. I call this a causally restricted interpretation of explanatory pluralism.

I believe that, in fear of the phantom of explanatory anarchism, 'causally restricted' pluralists may have granted too much to mechanists. I have indicated that the ontic interpretation of the idea of a mosaic unity is fully compatible with the proliferation of causal models about a target system. Part and parcel of the idea of a *mosaic* unity is that the findings

from different scientific fields are used, like tiles in a mosaic, to shape the SPM for a given phenomenon. Scientific fields are thought as groups of researchers related by central problems, experimental techniques, theoretical vocabularies and background assumptions (Craver, 2007; Darden & Maull, 1977). Some examples of neuroscientific fields are molecular neuroscience, molecular genetics, neurophysiology, neuroimaging, mathematical analysis, computational modeling and experimental psychology (Piccinini & Craver, 2011).

From the mechanistic perspective, scientific fields are not bounded by intertheoretical reductive links, as in classical reductionism (Nagel, 1961; Oppenheim & Putnam, 1958). Mechanists accept that fields are autonomous to the extent that each is allowed to choose which phenomena to explain, which experimental designs to apply, which conceptual resources to adopt, and the precise way in which they are constrained by scientific evidence from adjacent fields (Piccinini & Craver, 2011). The ability of scientific fields to provide novel constraints on the SPM for a given phenomenon is grounded on their relative autonomy. Craver (2007, p. 231) adopts the perspective metaphor when he asserts that "because different fields approach problems from different perspectives, using different assumptions and techniques, the evidence they provide makes mechanistic explanations robust."

Mechanism encourages the proliferation of causal models about the mechanism targeted for collective explanation, since that proliferation contributes to the robustness of a mechanistic explanation. Causally restricted pluralism and ontic mechanism seem to complement each other. The upshot of sections 2 and 3 is that no matter how many modeling strategies you may find in neuroscience, no matter how different they may seem from paradigmatic mechanistic explanations, if those strategies are construed as providing causal explanations, then they can be interpreted as tiles in the mosaic unity of a mechanistic research program. In section 4, I explore a *liberalized interpretation of explanatory pluralism*, i.e. one according to which (i) there are models in neuroscience that contribute to the collective explanation of some phenomenon φ but (ii) those models do not set any causal constraint on the SPM for φ.

Liberalized explanatory pluralism is an alternative to the central claim of mechanists, namely, that a model contributes to a collective explanation in neuroscience to the extent that it sets causal constraints on the target mechanism. A model that did not causally restrict the SPM would make no difference from a mechanistic point of view. A mechanist would say that such a model is, at most, a purely phenomenal model, a mere description of the phenomenon that does not carry any explanatory weight (Craver, 2006, 2007). My contention is that a liberalized interpretation of explanatory pluralism is the most reasonable approach to the nature of efficient coding explanation in computational neuroscience.

4. Efficient Coding Explanation

Chirimuuta (2014) argues that efficient coding explanation in computational neuroscience cannot be assimilated into the mechanist framework. Efficient coding explanations are provided by what Dayan and Abbott (2005, p. xiii) identify as "interpretative models." In their textbook of mathematical neuroscience, they distinguish between descriptive, mechanistic and interpretative models. Descriptive models characterize what neurons and neural circuits do, so they are very similar to what mechanists call 'phenomenal models:' their primary purpose is to describe phenomena, not to explain them. Dayan's and Abbott's 'mechanistic models' do what they are supposed to do according to the mechanistic philosophy: they explain how the nervous system operates, integrating constraints from multiple levels of mechanisms. Finally, interpretative models "use computational and information-theoretic principles to explore the behavioral and cognitive significance of various aspects of nervous systems function" (Dayan & Abbott, 2005, p. xiii). Interpretative models are purported to explain why nervous systems operate as they do, under the assumption that they are suited to the tasks they must carry out (Clatworthy, Chirimuuta, Lauritzen, & Tolhurst, 2003; see also Marr, 1982).

Chirimuuta (2014, p. 127) mentions that the kind of explanation interpretative models provide "bears interesting similarities with evolutionary and optimality explanations elsewhere in biology". The most conspicuous feature of optimality explanation is the use of mathematical techniques from the Optimization Theory framework. Optimality models represent and solve optimality problems, i.e., problems where the values of a given objective function on the set of possible solutions are to be maximized or minimized over a given constraint set (Sundaram, 1996). The objective function associates to each element of the set of possible solutions an element belonging to a totally ordered set of costs or values (Rosen, 1967). Crucially, optimality models

identify constraints on the set of possible solutions and tradeoffs among different costs or values to be achieved by the possible solutions. The problem of finding the optimal solution is then the problem of finding that solution corresponding to the minimum or the maximum value given the constraints and tradeoffs identified by the model.

Chirimuuta's (2014) case study of efficient coding explanation is Carandini's and Heeger's (2012) renaissance of Heeger's (1992) normalization model of simple cell response properties. Heeger's normalization model was proposed to explain non-linear properties of neurons in the primary visual cortex (are V1 in primates). On the one hand, excitation of cortical cells is highly specific to contrast independent features of a visual stimulus: cells are selective to orientations, spatial frequency and direction of motion. On the other hand, cortical cells have a limited dynamic range: their response is saturated by high contrasts. How is it possible for response ratios to be independent of stimulus contrast, in the face of response saturation? The original idea of Heeger (1992) was that each simple cell receives linear excitatory input from the lateral geniculate nucleus and it also receives inhibitory input from nearby neurons in the striate cortex. Therefore, these systems operate a divisive normalization: they compute a ratio between the response of an individual neuron and the summed activity of a pool of neurons.

Recently, Carandini and Heeger (2012, p. 51) have reinterpreted normalization as a canonical neural computation, that is, one of many "standard computational modules that apply the same fundamental operations in a variety of contexts." Other examples of canonical neural computations are exponentiation, linear filtering and gain control. These "recurring building blocks" (in terms of Weiskopf, 2011, p. 249) of cognitive systems have been reported to operate in several sensory modalities and anatomical regions, from auditory cortex to areas correlated with visual attention. Normalization may have a different function in each of these regions (e.g., discrimination amongst stimuli or redundancy reduction) but in each case the same computation is performed, namely, "dividing the output response of a neuron by a term that relates to the average firing rate of nearby neurons" (Chirimuuta 2014, p. 138). Furthermore, normalization considered as a neural computation is implemented by different biophysical mechanisms (such as synaptic suppression or shunting inhibition) in different brain regions. Thus, crucially, the phenomenon targeted by this efficient coding explanation seems to be multiply realized. The description of normalization that is relevant for efficient coding explanation does not demand a characterization of the biophysical mechanisms that implement the computation.

Why is normalization so widespread? Caradini's and Heeger's (2012) answer this question by providing an efficient coding explanation of normalization. Chirimuuta (2014, p. 144) argues that efficient coding explanations take an observed behavior and formulate an explanatory hypothesis about its "functional utility". In particular, Carandini's and Heeger's (2012) can be interpreted as arguing that "[normalization is so widespread] because for many instances of neural processing individual neurons are able to transmit more information if their firing rate is suppressed by the population average firing rate" (Chirimuuta 2014, p. 1430). This exemplar of efficient coding explanation is strongly analogous to examples of optimality explanation in elsewhere in biology. An optimality model identifies, first, a set of variables that describe the target system, called design variables. Different values for the variables produce different designs of the system. Carandini and Heeger (2012) seem to identify the computations a system may perform (linear filtering, thresholding, normalization etc.) as design variables. Second, an optimality model identifies an objective function for the optimum design problem, which needs to be maximized or minimized depending on the problem requirements. Carandini and Heeger (2012) take the maximization of information transmission as the optimization criterion. Third, an optimality model specifies certain restrictions or requirements placed on a design, called design constraints. Crucially, an optimality model may also specify certain tradeoffs among different costs or values. Carandini and Heeger (2012) take the dynamic range of cortical cells as a design constraint. Finally, an optimality model identifies which values of the design variables optimize the criterion of the model in light of the design constraints. Chirimuuta (2014, p. 146) summarizes this explanatory pattern as follows:

> The general strategy is to work from information theoretic fist principles to build a model of a hypothetical system which would maximize information transmission of the sort required by the brain area in question. Then one sees how the hypothetical optimal and real system line up with respect to neuronal response properties and other features. If there are similarities in the properties compared, we have an explanation of why the brain area has those properties.

What is the nature of efficient coding explanation? I have mentioned that efficient coding explanation is usually provided by interpretative models. Chirimuuta (2014) holds that most interpretative models are *minimal* models, i.e. models that highlight only a subset of the target system's features. More specifically, Chirimuuta differentiates between two kinds of minimal models. On the one side, there are B-minimal models. "B" is for Batterman, who introduced this kind of models in his (2002). B-minimal models typically define a 'universality class'. Two general characteristics of universality classes are the following: (i) "the details of the system (those details that would feature in a complete causal-mechanical explanation of the system's behavior) are largely irrelevant for describing the behavior of interest"; and (ii) "many different systems with completely different 'micro' details will exhibit identical behavior" (Batterman 2002, p. 13). Since B-minimal models omit most causal/mechanical information about the target system, they do not seem to impose any causal restriction on the SPM for the phenomenon in question. On the other side, there are A-minimal models. A-minimal models "include only the core causal factors which give rise to a phenomenon", that is, they contain "only those factors that make a difference to the occurrence and essential character of the phenomenon in question" (Weisberg 2007, p. 642). A-minimal models seem to be straightforwardly causal, thus they constrain in that way the SPM for the phenomenon in question. Are interpretative models more akin to A-minimal or to B-minimal models?

Chirimuuta (2014) somehow bypasses this question and asserts that interpretative models constitute a *sui generis* kind of model, namely, an I-minimal model. I-minimal models have two main features: (I) they "ignore biophysical specifics in order to describe the information processing capacity of a neuron or neuronal population"; and (II) "they figure in computational or information-theoretic explanations of why the neurons should behave in ways described by the model" (Chirimuuta 2014, p. 143). She correctly cautions that any attempt to assimilate I-minimal models with causal-mechanical models would obliterate the defining characteristics of efficient coding explanation. However, she concludes in 2014 that efficient coding explanations are causal after all. The reason, which may be appealing to many philosophers, is that efficient coding explanations "delineate a set of counterfactual dependences between input to the system (e.g. sensory information) and/or system requirements (e.g. task for which information is needed) and the computational properties of the system" (Chirimuuta 2014, p. 146). To the extent that interpretative models are able to address this kind of 'what-if-things-had-been-different questions' (or *w-questions;* Woodward, 2003, p. 221), they provide causal explanations. On this point I disagree. Chirimuuta (forthcoming) has changed her mind about the nature of efficient coding explanation in computational neuroscience. She argues that interpretative models do address w-questions, but they do not provide causal explanations, because the relevant counterfactuals that answer those questions do not describe ideal interventions on the target system.

In the first place, the capacity of a model to provide answers to w-questions by exhibiting counterfactual dependencies may not be grounded on the identification of causal relations within the target system. Lange (2013) claims, for example, that there are distinctively mathematical explanations in the empirical sciences that can address w-questions but not in virtue of describing the explanandum's causes. Pincock (2012) provides a particularly beautiful example of mathematical explanation in the empirical sciences, reviewed by Lange (2013, p. 488): "Why has no one ever succeeded (or why did a given person on a given occasion not succeed) in crossing the bridges of Königsberg exactly once?" (Figure 1).

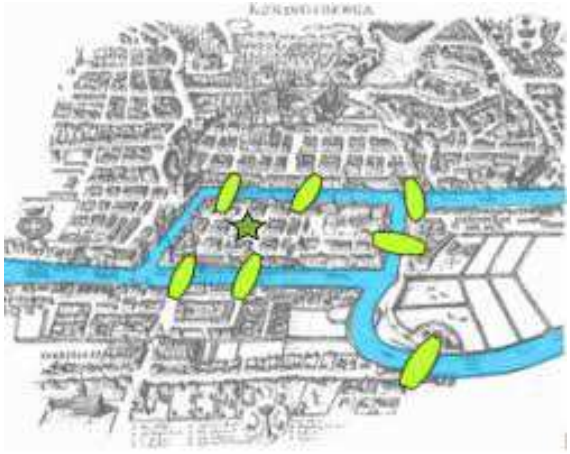

Figure 1
The bridges of Königsberg in Euler's time

Euler's mathematical explanation is that it is not the case that either vertex or every vertex but two is touched by an even number of edges. Although this explanation is clearly mathematical, it allows us to answer some w-questions, such as these: *had John's attempt to cross all the bridges of Königsberg exactly one begun on vertex D, he would have failed anyway*,

or *were there one more bridge and John's attempt to cross all the bridges at once would have been successful.*

In the second place, it is not the case that the information-theoretic constraints and tradeoffs that appear in efficient coding explanations are causal factors actually involved in the production of the explanandum phenomenon. Just like constraints and tradeoffs in optimality modeling elsewhere in biology, the information-theoretic constraints and tradeoffs of efficient coding explanations are not the sort of entities that participate in causal relations. Let us consider, for example, Harris's and Wolpert's (2006) model of saccade trajectories. They propose that saccade trajectories follow a stereotyped sequence because signal-dependent noise imposes a compromise between the speed and the accuracy of an eye movement and that the stereotyped sequence observed optimizes a tradeoff between the accuracy and duration of the movement. From an ontological point of view, tradeoffs like this are not events, nor causal properties within the system. Thus, the viability conditions that the speed-accuracy tradeoff imposes on the target system are not causes of the eye movement.

In the third place, let us consider the link that Chirimuuta (2014) establishes between efficient coding explanations and Mayr's (1961) ultimate causal explanations. The latter does not describe "a causal path leading to any current instantiation of the behavior or feature and so can easily be distinguished from local mechanistic explanations" (Chirimuuta 2014, p. 147). The explanatory information an interpretative model provides concerns the synchronic mathematical dependencies between abstract information-theoretic tradeoffs and computations. Chirimuuta (2014, p. 142) claims that "the use of 'normalization' in neuroscience *retains much of its original mathematical-engineering sense.* It indicates a mathematical operation —a computation— not a biological mechanism". These models do not attempt to represent the causal factors that produce or realize the phenomenon.

If efficient coding explanations are not causal, how do they contribute to collective explanations in neuroscience? This is a very important and thorny matter, so I will only make some preliminary suggestions. I hold that most interpretative models are explanatory because they are general. Of course, mechanist philosophers are well aware of generality as an explanatory virtue of models. For example, Craver (2009, p. 588) accepts that computational models of the hippocampus, despite being abstract with regard to almost every neurobiological detail, can provide a genuine explanatory payoff relative to other, more concrete, scientific models: "For some purposes (such as building an abstract computational model) generality is more important." However, efficient coding explanation (and optimality explanation in general) exhibits a kind of generality that is not identical to the kind of abstractive generality that characterizes mechanistic explanations.

'Generality' roughly refers to the number of target systems that a particular model or set of models applies to (Weisberg, 2007). This notion is ambiguous, since it entangles two different 'components' of generality together: A-generality and P-generality. A-generality corresponds to the number of target systems the model actually captures: P-generality is the number of possible, but not necessarily actual, target systems it applies to (Weisberg, 2007). P-generality is the kind of generality that is often thought to be associated with explanatory power. I would like to add that there is a frequently overlooked distinction between two kinds of generality: mechanism-bounded generality (or M-bounded generality), on the one side, and non-mechanistically bounded generality, or 'unbounded generality' (U-generality), on the other. More abstract models of a target system, to the extent that they set some causal constraints on the SPM for a given phenomenon, are relatively more 'general' than concrete, detailed models of the target system. The exhibit high M-bounded generality. The degree of abstraction of these censored causal explanations correlates with their M-bounded generality, because a model may be applicable to more instances of the mechanism as it represents relatively less causal components. My contention is that efficient coding explanations exhibit a distinct kind of generality, namely, U-generality. A mechanism sketch is more general than more complete models of the mechanism in question, but is not necessarily more general than other models which represent features that range across several mechanisms. Interpretative model in computational neuroscience are U-general in that the design features they represent may apply across several mechanisms in the brain.

It is fair to acknowledge that Boone and Piccinini (2016a) recognize *en passant* that some analyses of neural computation or information processing focus only on the information content and on the efficiency of a neural code, leading explanation outside the multilevel mechanistic framework. I think that it is a relevant datum for the debate between mechanists

and explanatory pluralists in the philosophy of neuroscience that the kind of explanation provided by interpretative models in computational neuroscience is distinct from mechanistic explanation. A mechanist philosopher may reply that the optimality approach is completely marginal or peripheral in neuroscience. It is important to remember that the center/periphery distinction is controversial. The tradition of optimality explanation in neuroscience has its roots in the works of Ramón y Cajal (1909). He first recognized that many aspects of brain organization can be accounted for by design features of the nervous system. Neuroscientists who advocate for canonical neural computations think that interpretative models in computational neuroscience are important tools in order to unlock the neural code and they even compare the search for canonical neural computations with the discovery of secondary structure in molecular biology (Caddick et al., 2009).

5. Conclusion

I have argued that, contrary to the ontic interpretation of Mechanism, there is at least one trend of scientific modeling in neuroscience, namely, efficient coding explanation in computational neuroscience, that contributes to collective explanations in the discipline without setting any causal constraint on the SPM for the relevant phenomena. It was not my aim in this paper to deny that the discovery and manipulation of mechanisms is central to neuroscientific practice. The motivation of a liberalized interpretation of explanatory pluralism is simply a growing awareness that there is a rainforest diversity of explanatory styles in neuroscience, and that many of them are on an equal footing.

Funding Information

I am grateful to The National Agency for Science and Technology Promotion (ANPCyT, PICT-2014-3422) and The National Research Council (CONICET) for support while writing this paper.

Acknowledgments

I am extremely grateful to Liza Skidelsky for her comments and advice. I also wish to thank Sabrina Haimovici, Nicolás Serrano, Abel Wajnerman, Mariela Destéfano, Fernanda Velázquez, and Sara Solcoff for extensive discussions on earlier drafts.

I declare that this article is my own original work and that any additional sources of information have been duly cited. This work has not been submitted for publication elsewhere. I do understand the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct and I unequivocally assert that this work conforms to that document.

References

Abney, D., Dale, R., Yoshimi, J., Kello, Ch., Tylén, K., & Fusaroli, R. (2014). Joint perceptual decision-making: a case study in explanatory pluralism. *Frontiers in Psychology, 5*(330), 1-12.

Batterman, R. (2002). *The devil in the details*. Oxford: Oxford University Press.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical perspective on cognitive neuroscience*. New York: Routledge.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A Mechanistic Alternative. *Studies in History and Philosophy of the Biological and Biomedical Sciences*, *36*, 421-441.

Boone, W., & Piccinini, G. (2016a). The cognitive neuroscience revolution. *Synthese, 193*(5), 1509-1534.

Boone, W., & Piccinini, G. (2016b). Mechanistic abstraction. *Philosophy of Science*, 83(5), 686-697.

Caddick, S., Carandini, M., Hausser, M., Martin, K., Priebe, N., Reynolds, … Yokoyama, C. (2009). Physiology: Mechanisms. In D. Heegger, E. Simoncelli, J. Reynolds, & M. Carandini (Eds.), *Canonical neural computation: a summary and a roadmap* (pp. 8-12). Recovered from: http://www.theswartzfoundation.org/docs/Canonical-Neural-Computation-April-2009.pdf

Carandini, M., & Heeger, D. (2012). Normalization as a canonical neural computation. *Nature Neuroscience*, *13*, 51-62.

Chirimuuta, M. (2014). Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience. *Synthese*, *191*, 127-153.

Chirimuuta, M. (forthcoming). Explanation in computational neuroscience: causal and non-causal.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge: MIT Press.

Clatworthy, P., Chirimuuta, M., Lauritzen, J., & Tolhurst, D. (2003). Coding of the contrasts in natural images by populations of neurons in primary visual cortex (V1). *Vision Research*, *43*, 1983-2001.

Craver, C. (2006). When mechanistic models explain. *Synthese, 28*(2), 141-163.

Craver, C. (2007). *Explaining the Brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon.

Craver, C. (2009). Mechanisms and natural kinds. *Philosophical Psychology, 22,* 575-594.

Craver, C. (2015). The ontic account of scientific explanation. In M. Kaiser, O. Scholz, D. Plenge, & A. Hüttemann (2015) *Explanation in the Special Sciences: The case of biology and history* (pp. 27-52). Dordretch:

Springer.
Dale, R., Dietrich, E., & Chemero, A. (2009). Explanatory Pluralism in Cognitive Science. *Cognitive Science, 33*, 739-742.
Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science, 44*, 43-64.
Dayan, P., & Abbott, L. (2005). *Theoretical Neuroscience: Computational and mathematical modeling of neural systems*. Cambridge: MIT Press.
Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London: New Left Books.
Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science,* 69(3), S342-S353.
Glennan, S. (2010). Mechanisms, Causes, and the Layered Model of the World. *Philosophy and Phenomenological Research*, *81*, 362–381.
Harris, C., & Wolpert, D. (2006). The main sequence of saccades optimizes speed-accuracy trade-off. *Biological Cybernetics, 95*(1), 25-29.
Heeger, D. (1992). Normalization of cell responses in the cat striate cortex. *Visual Neuroscience, 9*, 181-197.
Krickel, B. (forthcoming). A regularist approach to mechanistic type-level explanation.
Lange, M. (2013). What makes a scientific explanation distinctively mathematical?. *British Journal for the Philosophy of Science, 64*, 485-511.
Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421-431.
Levy, A. (2013). What was Hodgkin and Huxley's Achievement. *British Journal for the Philosophy of Science, 65*, 469-492.
Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 57*, 1-25.
Marr, D. (1982). *Vision.* San Francisco, CA: Freeman Press.
Mayr, E. (1961). Cause and effect in biology. *Science, 134*, 1501-1506.
Nagel, E. (1961). *The structure of science. Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace and World, Inc.
Oppenheim, O., & Putnam, H. (1958). Unity of Science as a Working Hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.). *Concepts, Theories and the Mind-Body Problem, Minnesota Studies in the Philosophy of Science II* (pp. 3-36). Minneapolis: University of Minnesota Press.
Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science*, *74*(4), 501-526.
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283-311.
Pincock, C. (2012). *Mathematical and Scientific Representation*. Oxford: Oxford University Press.
Ramón y Cajal, S. (1909). *Histology of the Nervous System of Man and Vertebrates*. Oxford: Oxford University Press.
Rosen, R. (1967). *Optimality Principles in Biology*. US: Springer.
Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.
Sundaram, R. (1996). *A First Course in Optimization Theory*. Cambridge: Cambridge University Press.
Thagard, P. (2003). Pathways to biomedical discovery. *Philosophy of science*, *70*(2), 235-254.
Weisberg, M. (2006). Forty years of 'The Strategy'. *Biology and Philosophy, 21*(5), 623-645.
Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy, 104*(2), 639-659.
Weiskopf, D. (2011). Models and mechanisms in psychological explanation. *Synthese 183*, 313-338.
Woodward, J. (2003). *Making things happen: a theory of causal explanation*. Oxford: Oxford University Press.
Wright, C. (2012). Mechanistic explanation without the ontic conception. *European Journal of Philosophy of Science, 2*(3), 375-394.
Wright, C., & Bechtel, W. (2007). Mechanisms and psychological explanation. In P. Thagard (Ed.), *Philosophy of Psychology and Cognitive Science* (pp. 31-79). New York: Elsevier.