Pluralistic Mechanism*

Abel Wajnerman Paz

Received: 09/08/2016
Final Version: 05/10/2016

BIBLID 0495-4548(2017)32:2p.161-175
DOI: 10.1387/theoria.16877

ABSTRACT: An argument recently proposed by Chirimuuta (2014) seems to motivate the rejection of the claims that every neurocognitive phenomenon can have a mechanistic explanation and that every neurocognitive explanation is mechanistic. In this paper, I focus on efficient coding models involving the so-called "canonical neural computations" and argue that although they imply some form of pluralism, they are compatible with two mechanistic generalizations: all neurocognitive explanations are (at least in part) mechanistic; and all neurocognitive phenomena that have an explanation have (at least) a purely mechanistic explanation.

Keywords: mechanism, pluralism, efficient coding, neural computation, neural coding.

RESUMEN: Un argumento recientemente propuesto por Chirimuuta (2014) parece motivar el rechazo de la tesis de que todo fenómeno neurocognitivo puede tener una explicación mecanicista y de la tesis de que toda explicación neurocognitiva es mecanicista. En este trabajo me centro en los modelos de codificación eficiente que involucran las llamadas "computaciones neuronales canónicas" y argumento que, aunque implican una forma de pluralismo, son compatibles con dos generalizaciones mecanicistas: todas las explicaciones neurocognitivas son (al menos en parte) mecanicistas; y todos los fenómenos neurocognitivos que tienen una explicación tienen (al menos) una explicación puramente mecanicista.

Palabras clave: mecanicismo, pluralismo, codificación eficiente, computación neuronal, codificación neuronal.

1. *Introduction*

Mechanism has evolved into a dominant perspective in the philosophy of neuroscience as it has proven to be a useful framework to account for the explanatory power of many models, ranging from molecular to behavioral neuroscience (e.g., Machamer *et al.,* 2000; Craver 2007; Bechtel 2008; Kaplan and Craver 2011; Piccinini and Craver 2011; and Boone and Piccinini 2015). David Kaplan (2011) argues that even explanatory models in "higher-level" branches of cognitive neuroscience, such as computational neuroscience, are mechanistic.

Chirimuuta (2014) claims, on the contrary, that the mechanistic approach cannot be universally applied to all neurocognitive phenomena. She affirms that relevant phenomena

[*] I thank Gualtiero Piccinini for a careful and very helpful review of earlier versions of this paper. I am grateful to the referees from Theoria for their insightful comments. My thanks also go to my advisor, Liza Skidelsky as well as to Sergio Barbieris, Sabrina Haimovici, Mariela Destéfano, Fernanda Velazquez, Nicolás Serrano, Magali La Rocca, Rodrigo Gonzáles Wilkens and Cristian Stábile for exhaustive discussions on many versions of this paper.

can only be explained by interpretative-minimal models, which provide efficient coding explanations. These are a form of the optimality explanations frequently employed in biology. They explain why a particular brain area or neural population has a given (e.g. computational) property, making reference to efficient coding principles.

In this paper, I focus on the minimal interpretative models that describe the so-called "canonical neural computations" (CNCs), and claim that although they imply a form of pluralism, they are compatible with some mechanistic generalizations about neurocognitive explanation. In section 2, I present mechanism, the characterization of normalization as a CNC, and interpretative minimal models. In section 3, I argue that the CNC-related phenomenon that a mechanistic model fails to explain (that is, widespread implementation of normalization) does not have an efficient coding explanation either. In turn, the phenomenon that can be explained by an efficient coding model (normalization in a particular brain area) can also be explained mechanistically. This implies a kind of pluralism that is compatible with a mechanistic claim: any neurocognitive phenomenon that has an explanation has (at least) a mechanistic explanation.

In section 4, I argue that many relevant optimality explanations in cognitive neuroscience presuppose mechanistic explanations. Efficient computation and efficient coding models of neural processing (two different kinds of optimality models) require determining not only that a given coding regime or computational process constitutes the optimal strategy to perform a given task, but also that the relevant neural population actually implements that strategy, i.e., it requires explaining how the population performs the relevant task. I argue that the answer to this "how" question constitutes a mechanistic explanation. This implies that efficient coding and efficient computation models are compatible with the claim that neurocognitive explanation is (at least in part) mechanistic.

2. *Mechanism and efficient coding*

To consider the implications of efficient coding explanation for a mechanistic proposal it is important to first clarify the notions of mechanism and mechanistic explanation. A mechanism can be defined as "[a] structure performing a function in virtue of its component parts, component operations, and their organization" (Bechtel and Abrahamsen 2005, 423). Mechanisms are active structures that perform functions, produce regularities, underlie capacities, or exhibit phenomena, doing so in virtue of the organized interaction among the mechanism's component parts and the processes or activities these parts carry out (Kaplan 2011).

According to mechanism, the explanatory force of the model for a given phenomenon depends on how accurately it describes the underlying mechanism. This commitment is expressed by Kaplan's "model-to-mechanism-mapping" (3M) condition (Kaplan 2011, 347):

> **(3M)** A model of a target phenomenon explains that phenomenon to the extent that (a) the variables in the model correspond to identifiable components, activities, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these (perhaps mathematical) variables in the model correspond to causal relations among the components of the target mechanism.

It is relevant to point out that Kaplan (2011) proposes 3M as a requirement for *constitutive* mechanistic explanation. Mechanists (e.g., Craver 2007) often distinguish between etiological models, which explain why a phenomenon occurs and constitutive models, which explain how it occurs. Etiological explanations cite the antecedent cause of a phenomenon, whereas constitutive explanations identify properties underlying the phenomenon within the relevant system. Both are kinds of mechanistic explanation. However, Kaplan`s proposal concerns constitutive explanation. He claims that explanatory models in computational neuroscience provide constitutive mechanistic explanations.

According to Chirimuuta, the explanation of the so-called "canonical neural computations" (CNCs) constitutes a counter-example to this thesis. She claims that the implementation of CNCs can only be explained by non-mechanistic computational models. I will focus here on one of the most studied CNCs: normalization. Heeger (1992) proposed the normalization model, a quantitative model of the response properties of simple cells in the primary visual cortex that respond to specific stimuli (bars) in specific orientations. Among other things, this model can explain the fact, implied by the phenomenon of cross-orientation suppression (COS) (Bonds 1989), that the response of simple cells is non-linear. COS occurs when a non-preferred stimulus (e.g., a horizontal bar) of a simple cell in V1 is presented at the same time as the preferred stimulus (vertical bar), and the response of the cell is smaller than its response to the preferred stimulus alone. This fact cannot be accounted for by the original model proposed by Hubel and Wiesel (1962). Heeger's idea is that each simple cell has a linear excitatory input from LGN but also an inhibitory input from adjacent neurons in the visual cortex. The relation between these inputs and their output is defined by the equation:

$$\bar{E}_i(t) = \frac{E_i(t)}{\sigma^2 + \sum_i E_i(t)}$$

Where $\bar{\text{E}}_\text{i}$ is the normalized response of a simple cell, t is the time, σ2 is a parameter that governs the contrast at which the neuron is saturated, and ΣE is the sum of inhibitory responses of all simple cells in the local population.

Carandini and Heeger (2012) argue that normalization is a CNC. CNCs are defined as standard computational modules that are implemented in many different systems and organisms. Other examples of CNC are linear filtering, recurrent amplification, associative learning, and exponentiation. They are presented as a toolbox of computational operations that the brain applies in different sensory modalities and anatomic regions and that can be described at a level of abstraction above their bio-physic implementation. Carandini and Heeger show that the normalization model has been successfully applied to the olfactory system in invertebrates, the retina (photoreceptors, bipolar cells, and retinal ganglion cells), V1 and superior visual areas (MT, V4, IT), the auditory cortex (A1), multisensory integration (MST), visual-motor control (LIP), and attention.

Chirimuuta (2014) maintains that the fact that normalization is a CNC can only have a non-mechanistic explanation. The widespread implementation of normalization cannot be explained mechanistically as the result of a similar underlying neural circuit, i.e., a canonical neural circuit. This is because normalization is known to be implemented by a variety of neural mechanisms. For example, shunting inhibition and synaptic suppression are different mechanisms that implement normalization in different brain regions. Chirimuuta considers that a more plausible alternative to explain canonical computation is the presence of equivalent demands in different systems. The convergence of computational processes can be accounted for by their efficacy in addressing common tasks, specifically, information-processing tasks. For example, Heeger (1992) proposed that contrast normalization in the primary visual cortex has an important role in maintaining the specific calibration of simple cells regarding a small range of stimulus orientations, independently of stimulus contrast. Given that maximizing stimulus selectivity is a requisite for reliable sensorial coding in different modalities, it can be expected that normalization will be implemented in non-visual areas and in invertebrates, although the bio-physical implementation is different.[1] Taking this idea, Chirimuuta characterizes an interpretative minimal model in the following way:

> **I-minimal Model**: Models which ignore biophysical specifics in order to describe the information-processing capacity of a neuron or neuronal population. They figure in computational or information-theoretic explanations of why the neurons should behave in ways described by the model.

Chirimuuta (2014) offers a brief characterization of the general explanatory strategy that these models involve. First, we use information theory to determine how information transmission of the sort required by a relevant brain area can be optimized. For example, a given neural system could require the reduction of redundancies, the maximization of its sensitivity to variations in the stimulus or the maximization of signal invariance with respect to some stimulus dimensions. Then, we build a model of a hypothetical computational operation which would optimize information transmission in the required way. Finally, we compare the optimal and real neural computation. If there are similarities, we have an explanation of why the brain area implements the relevant computation.

Chirimuuta points out that this explanatory pattern is similar to the one we find in optimality explanations in biology. The idea that optimality explanations imply a form of pluralism regarding explanation has been defended outside computational neuroscience. For example, Rice (2015) maintains that optimality models can provide non-causal (and, hence, non-mechanistic) explanations in biology. Rice (2015) offers a succinct characterization of optimality models that can be useful to understand efficient coding explanations.

[1] Chirimuuta (2014) characterizes the relation between normalization and its different underlying mechanisms as one of multiple realization. The notion of multiple realization has been questioned by some mechanists (e.g. Bechtel and Mundale 1999 and Milkowski 2013). If one rejects the notion, then it seems that there could not be *canonical* computation (the same computation implemented in different neural systems) if there is no canonical circuit or underlying mechanism (that is, if each instance of that computation is not realized by the same mechanism). Different mechanisms should be taken to realize a different computation. If this is the case, then there is no phenomenon (such as the widespread implementation of normalization) that mechanistic models fail to explain. However, I will not criticize Chirimuuta on these grounds. The main reason is that not all mechanists reject multiple realization (for instance, Piccinini and Maley 2014 propose a mechanistic approach to this notion). I thank an anonymous referee for pointing out the relevance of addressing the issue of multiple realization within mechanism.

These models result from the application of a mathematical technique called "optimization theory." This is a technique that can determine what values of some control variable(s) will optimize the value of some design variable(s) given a set of tradeoffs and constraints. An optimality model specifies what is called a "strategy set," the set of possible strategies defined (at least in part) by different sets of values of the control variables; a "currency" (the designs variables to be optimized); and an optimization criterion (what it means to optimize the design variables). With these elements, it specifies what is known as an "objective function," which connects each member of the strategy set to values of the design variable(s) to be optimized. The function determines an optimal strategy, which is the one that optimizes the criterion in light of the relevant constraints and tradeoffs. An optimality model explains the current state of a system by showing that it implements a strategy that optimizes the relevant criterion.

This framework can now be employed to offer a more detailed characterization of the efficient coding explanation of a CNC. In the first place, a currency (i.e., a design variable *i* to be optimized) is defined. In our example, the relevant neural systems require the maximization of stimulus selectivity (maximization of invariance regarding stimulus contrast). In the second place, a strategy set is defined by different computational operations (such as linear filtering, normalization, exponentiation, etc.) which can contribute to *i* optimization. With these elements, we specify an objective function that assigns a value of *i* to each computational strategy. The optimal strategy will be the computational process *c* that is correlated with the optimal value of *i*. The model explains the presence of *c* in a given brain area by showing that it is the optimal strategy in this sense.

One could object that if these models explain *why* a brain area implements a given computation, then they provide etiological explanations. This is, as mentioned above, a kind of mechanistic explanation. Chirimuuta points out that although efficient coding explanations need a background assumption that there is *some* set of processes at work which have a tendency to optimize solutions to coding problems (e.g. Laughlin 2001), they are not required to provide a characterization of these processes. They are not committed to *specific* processes of this kind. This idea raises a further concern. If this general etiological assumption plays no significant role (that is, a role relevant enough to consider that the models are etiological), one could wonder why efficient coding models are explanatory. Chirimuuta has an interesting answer to this question.

She points out that these models are explanatory according to a criterion accepted by mechanists themselves. Some mechanists (e.g. Kaplan 2011; Kaplan and Craver 2011) endorse Woodward's idea (e.g. Woodward 2003) that the explanatory power of a model is given by its ability to address what-if-things-had-been-different questions or "w-questions." Woodward considers that in order to address these questions, "a model must describe the conditions that 'make a difference' to the explanandum in the sense that changes in these factors lead to changes in the explanandum" (Woodward, forthcoming, p. 5). Chirimuuta considers that efficient coding models can address questions of this kind without making reference to etiology. For instance, the normalization model implies that if the task that the relevant neural system needs to perform and/or its sensory input were different, then the underlying computations would also be different. She offers the example of a study presented by Wainwright *et al.* (2001) which shows that the normalization parameters are adjusted by variations in the statistics of recent visual input.

I consider that this also tells us why the general etiological assumption does not make the explanation etiological. It is clear that this assumption contributes very little to addressing the relevant w-questions. It is the application of optimization theory what carries the more significant explanatory weight. The assumption only implies that the actual strategy will be close to the optimal, but it says nothing about how it would change if other relevant features of the system where modified, or how the system would change if the strategy was different. We use optimization theory to determine what (non-optimal) values the design variables would have if the strategy was different, how the optimal values for design variables would change if the constraints where different or what would be the optimal strategy if the design variables where different.

A second objection could be that if efficient coding models explain *why* a given event occurs and not *how* it occurs, then it seems that a constitutive mechanistic model fails to explain this "*why*-phenomenon" simply because this is not the kind of phenomenon that it is supposed to explain. As mentioned, etiological explanations are the ones that are supposed to address these why-questions. The fact that a *why*-phenomenon is explained by something that is not a constitutive mechanistic model does not seem problematic for mechanism regarding constitutive explanation (which is the kind of mechanism that Kaplan and Chirimuuta discuss). I will address this issue in the following section. However, I will not deny that efficient coding models imply some form of pluralism. My aim is to evaluate whether and in which way this pluralism limits the application of the mechanistic approach.

3. *Different perspectives on a single phenomenon*

As we saw, Chirimuuta affirms that there is a question to which only an efficient coding model provides an answer. This does not mean that this is not the type of question for which a mechanistic constitutive model *could* be proposed. This is a relevant point. Kaiser and Krickel (2016) suggest that etiological and constitutive mechanistic models explain different kinds of phenomena. Even when both aim to explain, in some sense, the same behavior (for instance, protein synthesis) the constitutive and etiological explanation will have different *explananda*. At first glance, it would seem that both models have the same explanandum: one can etiologically explain protein synthesis by describing how a certain sequence of causes leads to the synthesis of a protein, or one can constitutively explain protein synthesis by referring to the components of a cell and describing how they act and interact such that the cell synthesizes proteins. The authors point out that "on a closer inspection, however, it turns out that what we are explaining is not the same phenomenon, but two different phenomena: the etiological MEx [mechanistic explanation] explains the end-result (there being a protein) and the constitutive MEx explains the process of protein synthesis (we want to know what happens at every step of protein synthesis)" (Kaiser and Krickel 2016, 8).

If we accept this point, it follows that constitutive mechanistic explanation does not apply to every neurocognitive phenomenon. This is not the right kind of explanation for the phenomena that, for instance, etiological models explain. If efficient coding explanations explain phenomena which (in this sense) mechanistic constitutive models are not supposed to explain, then they would not be problematic for mechanism. This would not imply a form of pluralism that was not already implied by etiological explanations. Furthermore, this is supported by the fact that efficient coding models address the same "why" questions addressed by etiological models. However, I will not discuss this point. In what follows I will concede, for the sake of Chirimuuta's argument, that there is a CNC-related phenomenon that could have both an efficient coding and a constitutive mechanistic explanation and determine the consequences of this assumption.

Chirimuuta (2014) considers that the widespread implementation of normalization, i.e., that many systems exhibit the behavior described by the normalization equation could have a mechanistic explanation if there were a canonical neural circuit, that is, if normalization was implemented by the same mechanism type in each of the relevant systems. However, we have seen that there is no canonical circuit and therefore the mechanistic explanation of this phenomenon is actually not available. In contrast, the phenomenon has an efficient coding explanation because, according to Chirimuuta, the demand for maximization of stimulus selectivity is as widespread as normalization. Nevertheless, there are reasons to doubt that there is such a canonical informational requirement. Carandini and Heeger (2012) mention a wide variety of informational demands in different systems and organisms whose optimization requires normalization. Informational demands associated with normalization seem to be at least as diverse as its underlying circuits.

There are at least six different applications of normalization. First, normalization is thought to be employed in some systems to maximize sensitivity. It can adjust the gain of neural responses to efficiently use the available dynamic range, thereby maximizing sensitivity to changes in input. Its implementation in light adaptation in the retina enables high sensitivity to subtle changes in visual features over a huge range of intensities. Second, normalization also contributes to achieving invariance with respect to some stimulus dimensions. For example, the antennal lobe of the fly is thought to achieve odorant recognition and discrimination regardless of concentration through normalization. Third, it also can contribute to decoding a distributed neural representation. For example, it is known that normalization contributes to decoding the distributed representation of visual motion in area MT. Fourth, normalization can also make the neural representations of different stimuli more readily discriminable by a linear classifier, that is, it can optimize stimuli discrimination. Fifth, normalization can cause a neuronal population to operate in two regimes, averaging the inputs when these are approximately equal and computing a winner-take-all competition (max-pooling, selecting the maximum of inputs). Normalization is thought to perform max-pooling, for example, in neural areas responsible for attentional modulation by selecting the neural sub-population with the largest response. Finally, it is known that normalization contributes to redundancy reduction. For example, it helps V1 to reduce redundancy by incrementing the statistical independence of its responses to natural images.

The fact that many different informational demands in different systems and species are satisfied by appealing to the same computational strategy implies that the efficient coding explanation is not better than the mechanistic one to account for the widespread or canonical character of the strategy. In the same way that there is not a one-to-one relation between neural circuits and computations, there is neither a one-to-one relation between computations and informational demands. This, of course, does not mean that the model is not explanatory at all. It only implies that the model cannot be used to explain why normalization is so widespread. Therefore, the efficient coding model for *this* phenomenon cannot be used to argue for pluralism regarding neurocognitive explanation.

We can employ this approach to explain normalization *in a particular system*. *Different* efficient coding models (models that have different design variables, i.e., that describe different informational demands) can be provided to explain the presence of normalization in different systems. However, we can also offer different mechanistic explanations for these phenomena. If the relevant phenomenon is no longer widespread implementation but merely implementation in a given system, there is no reason to prefer an efficient coding approach to CNCs over a mechanistic one. The efficient coding and mechanistic explanations of this phenomenon are equally good.[2]

This implies a form of pluralism regarding neurocognitive explanation that does not limit the range of applications of the mechanistic approach. As we saw earlier, Chirimuuta's argument implies that there are phenomena which are only explained by non-mechanistic models. This implies a form of pluralism according to which different neurocognitive phenomena require different kinds of explanations. I have argued that, at least regarding CNCs, this is not the case. Both efficient coding and mechanistic approaches are adequate for *the same CNC-related phenomenon*. This implies a more radical form of pluralism because it allows different perspectives on a single phenomenon. However, it is also compatible with a general mechanistic claim about neurocognitive explanation. Although we cannot affirm that only mechanistic models are explanatory, we can affirm that every neurocognitive phenomenon that has an explanation has (at least) a mechanistic explanation.

4. Explaining how before why

In what follows, I argue that there is a general conceptual relation between mechanistic and optimality explanations in cognitive neuroscience, and that this relation implies that pluralism is compatible with a different mechanistic generalization. We saw that I-minimal models exploit informational and computational properties of neural processing to explain why a given neuron or neural population perform a given task in a given manner. They explain this fact by mathematically determining that the *actual* strategy constitutes the optimal strategy to perform the task. This means that the mathematical model that determines which is the optimal strategy is only part of the explanation. The explanation only works when we can also show that the optimal strategy is the one actually employed by the studied system to perform the relevant task, i.e., we need to determine *how* the system performs

[2] I mentioned that Kaiser and Krickel (2016) distinguish between the phenomena that etiological and constitutive mechanistic models explain. They also argue mechanistic models explain a specific kind of phenomenon. What we can call a "mechanistic phenomenon" is a behavior, activity, or process (what they call an "ocurrent") performed by an object that is different from the relevant mechanism underling this behavior. It is important to point out that the phenomenon we are considering can be described as a mechanistic phenomenon in this sense. Regarding normalization, we can indeed distinguish between an activity (normalization, i.e., divisive inhibition), an object (a neural circuit comprising at least two input neurons —the driving and the modulatory inputs— and an output neuron), and a mechanism (e.g., shunting inhibition, defined by the location of the axon terminals of the modulatory input [Blomfield 1974] and their shunting effect). I thank an anonymous referee for pointing out the relevance of relating normalization with this account of mechanistic phenomena.

the task. In the case we have been considering, we must determine not only that normalization is the optimal strategy to maximize stimulus selectivity, but also that the relevant neural population actually implements this strategy to optimize selectivity.

According to Chirimuuta, the answer to this "how" question is not a mechanistic explanation. She points out that, from a mechanistic perspective, the normalization equation cannot be considered a fully explanatory model but rather a mechanism sketch. A mechanism sketch is a model that omits details about the underlying mechanism of a phenomenon that are not yet known (Machamer *et al.,* 2000). The normalization model gives a quantitatively accurate prediction of COS and numerous other phenomena (Heeger 1992) describing the suppressive effect (ΣE) of the relevant inhibitory mechanism in a very schematic way. However, I consider that the model cannot be considered a sketch. Unlike sketches, the omission of information in this model is not due to an imprecise knowledge of the relevant mechanism. We have seen that mechanisms implementing normalization, such as shunting inhibition or synaptic depression, are well known. Furthermore, the model has the features associated with abstract mechanistic models (what Chirimuuta calls "A-minimal models"). The core idea behind these models is that they must only describe the aspects of a mechanism that are difference makers for a relevant phenomenon. These difference makers are features that cannot be changed or replaced without modifying the behavior of the system (Levy and Bechtel 2013).

The normalization model can be considered an A-minimal model, for instance, for the phenomenon of cross-orientation suppression (COS). The information omitted from the model is not about difference makers for COS. Normalization can produce the non-linear response of a neural population required by COS even if it is implemented by the different mechanisms mentioned earlier. On the other hand, if divisive normalization did not affect the response of simple cells in V1, then these would not have the relevant non-linear properties. As I mentioned, these properties are not predicted by the model proposed by Hubel and Wiesel, which did not include divisive normalization. This means that the normalization equation omits non-difference makers and includes difference makers regarding COS. It can be considered a minimal mechanistic model.

This implies that the answer to the "how" question required by the efficient coding explanation of normalization constitutes a mechanistic explanation. It is not only the case that, as I argued above, normalization can have both optimality and mechanistic explanations. The optimality explanation of normalization has a mechanistic explanation as a constitutive part. This makes the former compatible with a mechanistic thesis that is stronger than the one defended in the previous section. It is not only compatible with the claim that all mechanistic phenomena that can be explained have a mechanistic explanation, but also with the thesis that all neurocognitive explanations are, at least in part, mechanistic.

It is important to emphasize that although CNCs are described by a limited set of neurocognitive models these have a widespread relevance for cognition. As we saw, *canonical* neural computations are defined as standard computational modules that perform the same operations in a wide variety of contexts. This means that providing an argument that mechanistic and optimality explanations are interrelated in the explanation of CNCs provides strong support for the mentioned mechanistic generalization.

To strengthen this point, I will show that the generalization is also supported by another group of optimality explanations that have at least the same broad relevance in cognitive neuroscience. As I mentioned above, Chirimuuta considers that the explanations that involve CNCs are efficient coding explanations. However, it is important to distinguish between neural computations and neural coding, and their different contributions to efficient information transmission. The models previously considered are concerned with how computations contribute to the optimization of information transmission. The normalization model describes a neural computation without specifying a coding regime (without specifying, for example, whether the neural signal is rate- or time-coded). There is a good reason for this. Neural computations are "coding regime independent," i.e., they can be performed by circuits that operate under different coding regimes. Specifically, divisive normalization can be performed by sustained rate-coded signals or sparse temporally correlated signals (Silver 2010). On the contrary, code specification is relevant for some of the underlying non-computational mechanisms since they can only operate under one specific coding regime. For example, changes in shunting inhibition, in concert with high levels of synaptic-input-dependent noise, synaptic short-term depression, and dendritic Na+ channels (which can produce a depolarizing after potential), can only control neural gain under sustained rate-coded signaling regimes since conductance changes produce additive shifts during temporally correlated signaling (Shu, Hasenstaub, Badoual, Bal, and McCormick 2003). Therefore, coding regime is not constitutive of neural computation but rather of its underlying (non-computational) mechanisms. The optimality explanation of efficient computation is not an explanation of efficient neural coding.

Detailed optimality models have been developed to show how neural coding contributes to efficient information transmission. Classical theoretical work hypothesized that optimizing information transmission is a driving force in the evolution of neural codes (Barlow 1959). Barlow (1969) introduced the idea that neural codes should minimize redundant information and maximize representational capacity (an idea then developed by, for example, Adelsberger, Mangan and Levy 1992; Foldiak 1990 and Redlich 1993). Later, considering that the brain is one of the metabolically most active organs of the body (Sokoloff 1989), Levy and Baxter (1996) claimed that neural coding must result from an optimal compromise between energy and informational efficiency. They determined that there is an optimum in the number of cells that should be active to encode a condition in order to reduce energy expenditure and that distributed coding gives a large reduction in the energy needed.

In this vein, Attwell and Laughlin (2001) explain neural coding by constraining some ideas from of Levy and Baxter (1996) through a detailed energy budget for brain signaling. Unlike Barlow (1969), Attwell and Laughlin do not take the representational capacity of a system (the number of encoded conditions) to be a design variable to be maximized but rather a constant that constrains the impact of different coding strategies of the system on energy consumption. The authors consider a system that must represent 100 different sensory or motor conditions. A purely local coding strategy is to represent each of the 100 conditions by 1 active cell to denote an occurring condition. Attwell and Laughlin estimate the energy expenditure of this coding regime by taking R to be the ATP (Adenosine Triphosphate, the molecule that carries the energy needed for neural signaling) usage per cell on the resting potential, and A the *extra* ATP usage per cell on active signaling (action potentials plus glutamatergic signaling). This implies that the total ATP used by the system to signal 1 of 100 conditions under this local coding regime would be 100R + A. When we begin to depart from this local coding regime towards a sparse one, an increase in energy efficiency is patent. If a condition is represented by the simultaneous firing of 2 cells (at the same rate, with the others not firing), only 15 neurons are needed to represent 100 conditions. This is given by the equation (which we can call the "capacity/components/code equation" or "3C equation") that relates representational capacity or number of conditions represented (R) with the number of cells or components of the system (n) and number of cells active to represent a condition (np):

$$R: n!/[(n - np)!\,(np)!]$$

In our case, 3C implies that 15!/(13! 2!) = 105. When we use this code, the energy expenditure is 15R + 2A. If R and A are equal (Attwell and Laughlin estimate that this is the case for neurons firing at 0.62 Hz), then this distributed representation gives a 6-fold reduction in energy usage for transmitting the same information. Similarly, if a condition is represented by 3 cells firing, 3C implies that only 10 cells are needed to represent 100 conditions (given that 10!/(7! 3!) = 120), and the energy expenditure is 10R + 3A, which (for R = A) is a further improvement of energy efficiency. Atwell and Laughlin point out that optimal neural coding is restricted not only by representational capacity but also by the temporal resolution required by the relevant informational task. This restriction is expressed by variations in firing rate. If the system needs a higher temporal resolution, the active cells must fire at a higher rate, so that when a new condition occurs, the switch in identity of the cells firing will become detectable earlier.

They claim that the sparseness of the optimal coding increases as the required firing rate increases. For signaling by active cells at 0.62 Hz, this optimum is broad, with 3 (of 10) or 4 (of 9) cells simultaneously active to optimally encode a condition. If active cells signal at 4 Hz, for which Attwell and Laughlin's calculations give A = 6.4R, then the optimum becomes sharper and has just 2 cells (of 15) simultaneously active to optimally encode a condition. Finally, if active cells signal with action potentials at 40 Hz, for which the budget implies A = 64R, then the optimum becomes sharper still.

It is clear that these progressively more precise characterizations of optimal coding can explain actual neural coding only if a given population actually implements the optimal strategy to represent a number of conditions. As we saw, the last step of an optimality explanation is determining whether the optimal hypothetical strategy lines up with the actual one. The explanation requires determining how the population actually represents the relevant condition. As with efficient computation, the answer to this "how" question requires a mechanistic model. A common approach to neural coding involves population analysis (Quian Quiroga and Panzeri 2013). It is usually assumed that to understand neural code we have to look for patterns in the combined activity of different neurons. There are two main strategies to analyze the activity of neural populations (Quian Quiroga and Panzeri 2009). Decoding algorithms can be used to reconstruct a given stimulus from the pattern of responses of a given neural population (Abbott 1994; Rieke *et al.,* 1997; Oram *et al.,* 1998; Pouget *et al.,* 2000; Dayan and Abbott 2001). Also, the concept mutual information can be employed to determine how much information (in bits) neurons carry about the stimuli (Deco and Obradovic 1997; Rieke *et al.,* 1997; Borst and Theunissen 1999; Dayan and Abbott 2001).

There can be distinguished three main steps for the analysis of multiple single-cell recordings (Quian Quiroga and Panzeri, 2009). The first is the extracellular recording of the activity of neurons with intracranial electrodes. The second is the discrimination of activity of single neurons from the continuously recorded data by means of spike detection and sorting algorithms. These algorithms make possible to identify a pattern of multiple spike trains. Lastly, this pattern is interpreted using decoding or information theory. The basic idea of these approaches is to quantify the amount of information in the spike trains about the stimulus (information theory) or to predict what stimulus produces the observed spike trains (decoding). Once we are able to make these predictions, we can then explore systematically which features of the spike trains carry the relevant information, i.e., which code the population implements. For example, with a population analysis we can establish whether the information about the observed stimulus is given by an increase in the firing of one of the neurons, or by a particular time firing pattern, or by a correlated firing of two neurons.

I consider that the result of population analysis with decoding or information theory are mechanistic models. They explain neural signaling by describing the relevant population as a mechanism, i.e., as the organized activity of components. The two approaches consider the information as a result of the activity of the population of neurons as a whole, and can determine how each member of the population (the components) and their interactions (their activities and causal organization) contribute to stimulus representation (Quian Quiroga and Panzeri 2013). For example, the model can determine how the ambiguity in the signal of a given neuron can be resolved by the activity of other neurons of the population. They can resolve this ambiguity, for example, by coordinating their relative time of firing to tag particularly salient events (Singer and Gray, 1995; Engel and Singer, 2001) or, alternatively, by having each neuron representing separately a particular stimulus or stimulus feature (Barlow *et al.,* 1964; Reich *et al.,* 2001; Quian Quiroga *et al.,* 2005).

These considerations imply that the modeling of neural coding (which constitutes part of efficient coding explanations) can be considered a mechanistic explanation of neural signaling. Given that neural codes (in the same way as neural canonical computations) constitute a pervasive aspect of neurocognitive processes, this case provides strong support to the thesis that neurocognitive explanations are, at least in part, mechanistic.

5. *Conclusion: Pluralism and Integration*

Efficient coding models have proven to be very useful tools to understand different aspects of neural processing. In the face of this fact, it is not possible to claim that (purely) mechanistic models are the only explanatory neurocognitive models. However, I tried to argue that the form of pluralism implied by efficient coding explanations is compatible with relevant mechanistic generalizations.

This pluralism has a further implication. I consider that it can be useful to understand neurocognitive integration. It is an important challenge for any approach to explanation in cognitive neuroscience to determine how different explanations in the field are integrated. Traditional cognitive science suffered from a strict division of labor between different explanatory strategies. Models at a functional or cognitive level (or, in Marr's terminology, a computational or algorithmic level), on the one side, and models at a neural, mechanistic or implementation level, on the other side, were considered distinct and autonomous from one another. In contrast, the recent development of cognitive neuroscience is gradually undermining this division. It is apparent that the computational and neural approaches to cognition are being increasingly connected (Boone and Piccinini 2015). Mechanism is a promising tool to characterize this integration. The different models can be connected if they are descriptions of different levels of the same mechanism. However, if efficient coding models are not (purely) mechanistic, perhaps we need a different, richer view on integration.

Chirimuuta (2014) suggests that we should understand neurocognitive integration in terms of a "perspectivist" approach. She considers that different kinds of explanations are complementary when they constitute different perspectives on a given system. She affirms that "the same system in neuroscience can be represented and modelled in a variety of different ways, depending on the particular purposes of the investigation" (p. 148). I consider that the idea of different perspectives on the same system is not informative enough to account for integration. Specifically, it does not say anything about how an efficient coding and a mechanistic model of a system could be connected. The mechanistic approach provides a characterization of the connection between different models. Although the notion of a mechanistic level is still controversial (e.g., Craver 2015), we know that different levels described by different mechanistic models of a system have some kind of compositional relation. To offer some insight into integration, the perspectivist view should provide a characterization of these relations.

An alternative strategy is to understand integration as something relative to a given phenomenon. Different explanations are integrated insofar as they explain the same phenomenon. According to this view, integration does not require any relation between the aspects described by different *explanans*. However, we have seen that Chirimuuta's considerations about the efficient coding explanation of CNCs do not support this kind of integration. She argues that for some phenomena (such as the widespread implementation of normalization) only the efficient coding perspective is adequate. On the contrary, my argument in section 3 supports this perspectivist view. Both mechanistic and efficient coding approaches are adequate perspectives on the implementation of normalization by a given brain area.

I consider that the argument provided in section 4 implies a different kind of integration. It suggests that at least some efficient coding and mechanistic explanations do not have the same explanandum. A mechanistic model can explain how a system computes or codes information and an efficient coding model can explain why that system computes or codes information in that way. Given this difference, these two explanations cannot be considered perspectives on the same phenomenon. However, they are connected by a conceptual relationship: explaining why (in the sense of providing an efficient coding or an efficient computation explanation) presupposes explaining how. This relationship determines a different form of integration between neurocognitive explanations. This implies that we should not be pluralists only about explanatory strategies, but also about their integration. Different kinds of explanations are connected in more than one way.

REFERENCES

Abbott, Larry. 1994. Decoding neuronal firing and modelling neural networks. *Q Rev Biophys* 27: 291-331.

Adelsberger-Mangan, Dawn and William Levy. 1992. Information maintenance and statistical dependence reduction in simple neural networks. *Biol. Cybern.* 67: 469-477.

Attwell, David, and Simon Laughlin. 2001. An Energy Budget for Signaling in the Grey Matter of the Brain. *Journal of Cerebral Blood Flow and Metabolism* 21:1133-1145.

Barlow, Horace. 1959. *Symposium on the Mechanization of Thought Processes*. H. M. Stationary, London, No. 10: 535-539.

Barlow, Horace., Mathew Hill and William Levick. 1964. Retinal ganglion cells responding selectively to direction and speed of image motion in the rabbit. *J Physiol* (London) 173: 377-407.

Barlow, Horace. 1969. Trigger features, adaptation and economy of impulses. In *Information Processing in the Nervous System*, edited by Nicholas Leibovic, 209-226. Springer-Verlag, New York.

Bechtel, William and Adele Abrahamsen. 2005. Mechanistic Explanation and the Nature-Nurture Controversy. *Bulletin d'Histoire Et d'pistmologie Des Sciences de La Vie* 12: 75-100.

Bechtel, William and Jennifer Mundale. 1999. Multiple Realizability Revisited: Linking Cognitive and Neural States. *Philosophy of Science* 66: 175-207.

Bechtel, William. 2008. *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Blomfield, Stephen. 1974. Arithmetical operations performed by nerve cells. *Brain Res*. 69: 115-124.

Bonds, A. B. 1989. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience* 2: 41-55.

Boone, Worth and Gualtiero Piccinini. 2015. The cognitive Neuroscience Revolution. *Synthese*. Published online, DOI: 10.1007/s11229-015-0783-4.

Borst, Alexander and Frédéric Theunissen. 1999. Information theory and neural coding. *Nat Neurosci* 2: 947-957.

Carandini, Mateo and David Heeger. 2012. Normalization as a canonical neural computation. *Nature Reviews Neuroscience* 13: 51-62.

Chirimuuta, Mazviita. 2014. Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese* 191(2): 127-154.

Craver, Carl. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

—. 2015. Levels. *OpenMIND project*.

Dayan, Peter and Larry Abbott. 2001. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, Massachusetts: MIT Press.

Deco, Gustavo and Dragan Obradovic. 1997. *An Information-Theoretic Approach to Neural Computing*. Berlin: Springer.

Engel, Andreas Karl and Wolf Singer. 2001. Temporal binding and the neural correlates of sensory awareness. *Trends Cogn Sci* 5: 16-25.

Foldiak, Peter. 1990. Forming sparse representations by local anti-hebbian learning. *Biol. Cybern*. 64: 165-170.

Heeger, David. 1992. Normalization of cell responses in the cat striate cortex. *Visual Neuroscience*, 9: 181-197.

Hubel, David Hunter and Torsten Nils Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160: 106-154.

Kaiser, Marie I. and Krickel, Beate. 2016. The Metaphysics of Constitutive Mechanistic Phenomena. *The British Journal for the Philosophy of Science.* doi:10.1093/bjps/axv058

Kaplan, David M. 2011. Explanation and description in computational neuroscience. *Synthese* 183(3), 339-373.

Kaplan, David M. and Carl Craver. 2011. The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science* 78: 601-627.

Laughlin, Simon. 2001. Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology* 11: 475-480.

Levy, William and Robert Baxter. 1996. Energy-efficient neural codes. *Neural Computation* 8: 531-543.

Levy, Arnon and William Bechtel. 2013. Abstraction and the Organization of Mechanisms. *Philosophy of Science* 80 (2):241-261.

Machamer, Peter, Lindley Darden, and Carl Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67 (1): 1-25.

Miłkowski, Marcin. 2013. *Explaining the Computational Mind*. Cambridge, Mass.: MIT Press.

Oram M. W., Foldiak P., Perrett D. I. and Sengpiel F. 1998. The Ideal Homunculus: Decoding neural population signals. *Trends Neurosci* 21:259-265.

Piccinini, Gualtiero and Carl Craver. 2011. Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183(3): 283-311.

Piccinini, Gualtiero and Corey Maley. 2014. The metaphysics of mind and the multiple sources of multiple realizability. In *New waves in the philosophy of mind*, edited by Mark Sprevak and Jesper Kallestrup, 125-152. New York: Palgrave Macmillan.

Pouget Alexandre, Peter Dayan and Richard Zemel. 2000. Information processing with population codes. *Nat Rev Neurosci* 1: 125-132.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., Fried, I. 2005. Invariant visual representation by single neurons in the human brain. *Nature* 435: 1102-1107.

Quian Quiroga Rodrigo and Stefano Panzeri. 2009. Extracting information from neural populations: Information theory and decoding approaches. *Nat Rev Neurosci* 10: 173-185.

Quian Quiroga, Rodrigo and Stefano Panzeri, 2013. *Principles of Neural Coding*, CRC press, Boca Raton, FL.

Redlich, Norman. 1993. Supervised factorial learning. *Neural Comp.* 5: 750-766.

Reich, Daniel, Ferenc Mechler, and Jonathan Victor. 2001. Independent and redundant information in nearby cortical neurons. *Science* 294: 2566-2568.

Rice, Collin. 2015. Moving Beyond Causes: Optimality Models and Scientific Explanation. *Noûs* 49 (3): 589-615.

Rieke F., D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek. 1997. *Spikes: Exploring the Neural Code.* Cambridge, MA: MIT Press.

Shu, Y., A. Hasenstaub, M. Badoual, T. Bal, and D. A. McCormick. 2003. Barrages of synaptic activity control the gain and sensitivity of cortical neurons. *J. Neurosci.* 23: 10388-10401.

Sokoloff, Lois. 1989. Circulation and energy metabolism of the brain. In *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects*, edited by G. J. Siegel,. W. Agranoff, R. W. Albers, and P. B. Molinoff,, 4th ed., 565-590. Raven Press, New York.

Silver, Angus. 2010. Neuronal arithmetic. *Nature Reviews Neuroscience* 11: 474-489.

Singer, Wolf and Gray, Charles. 1995. Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci* 18:555-586.

Wainwright, M. J., O. Schwartz, and E. Simoncelli. 2001. Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons. In *Statistical theories of the brain,* edited by R. Rao, B. Olshausen & M. Lewicki. Cambridge, MA: MIT Press.

Woodward, James. 2003. M*aking Things Happen*, New York: Oxford University Press.

— (forthcoming). Explanation in neurobiology: An interventionist perspective. In *Integrating Psychology and Neuroscience: Prospects and Problems,* edited by David Kaplan. Oxford: Oxford University Press.

Abel Wajnerman Paz has a PhD from the University of Buenos Aires. His doctoral research project was funded by CONICET scholarships (2010-2013 and 2013-2015). His present research project is focused on the relation between mechanistic and computational models in cognitive neuroscience and is funded by a CONICET postdoctoral scholarship (2015-2017). He is a member of the Cognition, Language and Perception Research Group from Buenos Aires, which carry out the research project PICT-2014-3422 funded by ANPCyT and is directed by Liza Skidelsky.

Address: Facultad de Filosofía y Letras de la Universidad de Buenos Aires, Instituto de Filosofía "Dr. Alejandro Korn". Puán 480, 4to. piso, of. 431 (1406), Buenos Aires, Argentina. E-mail: abelwajnerman@gmail.com