Abstract Mechanisms and Neural Computation

Abel Wajnerman Paz

CONICET (National Scientific and Technical Research Council)
abelwajnerman@gmail.com

A characterization of computation and computational explanation is important in accounting for the explanatory power of many models in cognitive neuroscience. Piccinini (2015) describes computational models as both abstract and mechanistic. This approach stands in contrast to a usual way of understanding mechanism which implies that explanation is impoverished by abstraction. I argue that in order to provide a useful account of computational explanation in cognitive neuroscience, Piccinini's proposal must be complemented by an abstraction criterion that fulfills two conditions: motivating abstractions enough to make a model computational and not motivating the omission of information that is constitutive of mechanistic explanation. These conditions are relevant because although there are computational and mechanistic descriptions of neural processes (Piccinini & Bahar 2013) mechanism must, as a normative theory, determine whether the abstractions that these models involve are well motivated. I argue that the abstraction criterion proposed by Levy and Bechtel (2013) is a promising candidate to fulfill these requirements. First, I show that this criterion can legitimize the omission from recently proposed neurocognitive models of all features that are non-computational according to Piccinini's approach (although it also motivates some modifications of his characterization of neural computation). Second, I argue that this criterion legitimizes those models only if we interpret them as including all the information constitutive of mechanistic explanation.

Keywords: *abstraction, mechanism, computational explanation, canonical neural computations, neural circuits*

***Journal of Cognitive Science* 17-1: 1-26, 2016**
Date submitted: 02/18/16 Date reviewed: 02/23/16
Date confirmed for publication: 03/03/16

1. Introduction

Piccinini (2015) develops a mechanistic characterization of computation. This approach implies that (at least some) computational explanations and specifically computational explanations in cognitive neuroscience can be mechanistic. According to Piccinini (2015), a constitutive feature of computational models is that they describe the relevant mechanism in an abstract manner. Both advocates (e.g., Kaplan) and critics (e. g., Chirimuuta 2014 and Haimovici 2013) of mechanism maintain that many computational models are abstract, but consider that mechanism implies that abstraction diminishes explanatory power (i.e., that more detailed non-computational descriptions are more explanatory). I agree with Piccinini that computational description is abstract (in the "medium independent" sense that he proposes, which is explained below) and that mechanism is compatible with the idea that abstract (and, specifically, computational) descriptions can be as explanatory as more detailed descriptions of a mechanism. However, I consider that in order to be useful for understanding computational models in cognitive neuroscience, Piccinini's approach must be complemented by an abstraction criterion that fulfills two conditions: motivating abstractions enough to make a model computational and not motivating the omission of information that is constitutive of mechanistic explanation.

The second condition is important because there is information which cannot be omitted without making the explanation non-mechanistic. I follow Levy and Bechtel (2013) in the idea that an explanation is mechanistic only if it describes causal organization, i.e., only if it describes the different causal contributions of the mechanism underlying a phenomenon and how these contributions are integrated. Models that omit any reference to causal organization do not constitute mechanistic explanations. An abstraction criterion that legitimizes *mechanistic* computational models must not motivate the omission of causal organization. The first condition is relevant because although there do, in fact, exist descriptions of neural processes that are abstract enough to count as computational according to Piccinini's notion (Piccinini and Bahar 2013) mechanism must, as a normative theory, determine whether the abstractions in these descriptions have an adequate epistemic motivation. Piccinini and Bahar (2013) argue that spike timing

and rate are often considered the vehicles of neural processing and that they possess the features constitutive of computational vehicles. But the mere facts that these descriptions are available and that they individuate relevant functional features of neural processing do not imply that the descriptions of neural processing that *only* make reference to these features are explanatory. In section 2, I elaborate on these two motivations for an abstraction criterion.

In sections 3 and 4, I argue that the criterion proposed by Levy and Bechtel (2013) ("LB") is a promising candidate to fulfill these requirements. Chirimuuta (2014) has recently considered some neurocognitive models that describe so-called "canonical neural computations" and argued that they are involved in explanations that cannot be characterized as mechanistic, in some kind of optimality explanation. She also maintains that these models are descriptions that are above a mechanistic level of abstraction and therefore cannot be employed to provide mechanistic explanations. In section 3, I show that LB can legitimize the exclusion from these models of all features that are non-computational according to Piccinini's approach. I argue that their explanatory power can be accounted for by this criterion; that LB can show why it is relevant to omit information about the neural circuits underlying computational operations. I also argue that LB motivates some modifications of Piccinini's characterization of neural computation. In particular, I show that neural coding regimes (that is, time-sensitive sparse coding or rate coding) are not part of neural computation, but rather of the underlying neural circuits.

In section 4, I argue that LB legitimizes the models that describe canonical neural computations only if we interpret them as including all the information constitutive of mechanistic explanation. LB cannot legitimize them if we interpret them as describing merely mathematical relations between the relevant variables while omitting any reference to their causal relations. The models can be considered explanatory only if they do not merely describe a mathematical equation, but rather a set of mathematically characterized causal relations; for example, if "division" is not interpreted literally as division, but as *divisive inhibition* of the activity of one component by the activity of another.

If this is correct, the models that describe canonical neural computations

can be considered both mechanistic and computational descriptions of neural processes. Although this argument concerns a limited set of neurocognitive models, its relevance can be enhanced once we realize that canonical computations are computational modules that are thought to operate in a wide variety of neural systems and are combined in different ways to perform different cognitive tasks. Canonical computations could provide a principled way or a unified language to understand neurocognitive function (Carandini & Heeger 2012). Although these models describe a limited set of neural operations, their relevance for cognition is likely to be very broad.

2. Computation and mechanism

Although there are more or less standard ways to characterize different types of computation, there is not widespread agreement regarding a general notion of computation. Piccinini (2015) proposes a notion that can be applied equally to the more general types of computation, i.e., a notion that describes the features shared by the most general kinds of computation (analog and digital). Generic computation is the processing of vehicles by a functional mechanism according to rules ([input × internal states/ output] mappings) that are sensitive only to differences between portions (spatiotemporal parts) of the vehicles. This feature of the rules that define computation implies that computational processes are abstract. When we define a computational process, we do not need to consider all of its specific physical properties. We can consider only the properties that are relevant according to computational rules. Given that the vehicles of concrete computation can be defined independently of the physical medium that implements them, Piccinini calls them (borrowing the term from Garson 2003) "medium independent." In other words, computational descriptions of concrete physical systems are sufficiently abstract so as to be considered medium independent (Piccinini 2015, pp. 121-2).

To put it another way, if the rule that defines a computation is sensitive only to differences between portions of the vehicles regarding specific dimensions of variation (if it is insensitive to any other physical property of the vehicles), then the vehicle is medium independent. Rules that define

computation are functions of state variables associated with certain degrees of freedom, which can be implemented in different physical media (for example, mechanical, electro-mechanical, electronic, magnetic, etc.). In contrast, typical physical processes, such as cooking, cleaning, or exploding, are defined in terms of specific physical processes that involve specific physical and chemical components. They are not medium independent.

A common way to understand mechanistic explanations is *prima facie* incompatible with the thesis that computational descriptions are at the same time abstract (insofar as they describe medium independent processes) and explanatory. To show why, I will briefly characterize 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, p. 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, pp. 346-7).

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-mechanism-mapping" (3M) condition (Kaplan 2011, p. 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.

Kaplan considers that mechanism has an additional commitment. It requires that the description of the mechanism be as complete as possible. Kaplan claims that the more precise and detailed the model of a phenomenon is, the better it explains. One can improve the quality of an explanation by including more mechanistic details in the model, e.g,

including additional variables to represent additional components of the mechanism. Chirimuuta (2014) calls this requirement "the More Details the Better" (MDB). A given model is more explanatory if it does not describe only the computational processes or algorithms underlying a cognitive capacity, but also all biological and bio-chemical processes that implement them.

In the face of the contrast between understanding computational models as providing abstract and fully explanatory descriptions and the idea that a model explains better if it omits the least possible detail, Haimovici (2013) poses a dilemma for Piccinini's notion of generic computation. We can develop it in the following way. Haimovici points out that a mechanistic model has to specify the causal structure of a system in terms of both structural and functional properties. However, generic computation barely appeals to structural properties, being mainly concerned with functional properties. This is what makes generic computation a notion that preserves the sense of "computation" in computer science. So, Piccinini's account faces a dilemma. If our notion of computation respects the mechanistic ideal of including all structural properties, the models we call "computational" are fully explanatory, but they are not computational in a relevant sense. If the notion of computation is defined mainly by functional properties, then it can be applied to all relevant computational systems, but computational models are not fully explanatory[1].

I consider that Haimovici's dilemma follows from the common way of understanding mechanism outlined above. For this reason, rejecting this form of mechanism is a crucial element in Piccinini's response. Piccinini (2015) makes two points. First, he argues that computational vehicles are individuated not only by their functional properties, but also by their structural properties. I consider that this point is not a problem for Haimovici, as she points out that computational descriptions *barely* appeal to structural properties, implying that although some are specified, many are omitted. Piccinini accepts that computational description is abstract in some sense (that it omits at least some structural properties). Given the MDB requirement, this is enough to trigger the dilemma. Second, Piccinini

[1] Haimovici (2013): 152 ss.

considers that mechanists need not endorse something like MDB. Piccinini rejects the idea that mechanistic explanations require the specification of structural and functional properties at all levels of organization. On the contrary, mechanistic explanations require the specification of all functional and structural properties at the relevant level of the mechanism. Omission of lower-level information is legitimate. We can be mechanists regarding computation and still preserve a characterization of computational models as abstract mechanistic descriptions[2].

Nevertheless, Piccinini does not specify the epistemic motivation behind the abstractions required by computational models, i.e., he does not provide the reason why we should explain some cognitive phenomena by appealing to medium independent descriptions instead of more concrete biophysical models. Therefore, I believe that we need an abstraction criterion to defend the idea that computational models are (fully) explanatory. Furthermore, I maintain that a criterion that legitimizes a mechanistic computational model must fulfill two conditions: motivating abstractions enough to make the model computational and not motivating the omission of information that is constitutive of mechanistic explanation. The second condition is important because there is information which cannot be omitted without making the explanation non-mechanistic. The first condition is relevant because although there are computational descriptions of neural processes (Piccinini and Bahar 2013), as a normative theory, mechanism must determine whether the abstractions in these models are well motivated[3]. In what follows, I address these two points.

First, there seem to be pieces of information that are constitutive of mechanistic explanation. A criterion that legitimizes a *mechanistic* computational model must not motivate the omission of this information.

---

[2] Ibid. pp. 124 and 125.

[3] Boone and Piccinini (unpublished) consider the relation between abstraction and mechanistic explanation, but they address a different (although related) question regarding this relation. They consider which forms of abstraction are involved in mechanistic explanation and what criteria motivate them. Here, I am concerned with determining whether an adequate criterion motivates, in some specific neurocognitive models, sufficient omissions for medium independence and prevents in those models the omission of information constitutive of mechanistic explanation.

Levy and Bechtel (2013) maintain that mechanistic explanations address *organized* systems. They argue that a system is organized with respect to a given behavior if different components of the system make different contributions to that behavior and the component's differential contributions are integrated (i.e., each component interacts in particular ways with a subset of the other components) (Levy & Bechtel 2013, p. 244). An explanation is mechanistic only if it describes organization, i.e., only if it describes the different contributions of the mechanism underlying a phenomenon and how these are integrated. This makes sense, given that causal organization is part of what constitutes a mechanism according to the definition explained above. If a model fails to refer to the constitutive features of a mechanism, there is no clear sense in which we can say that it provides a mechanistic explanation. So, an abstraction criterion for a mechanistic computational model must not motivate the omission of causal organization.

I will argue that the criterion proposed by Levy and Bechtel (2013) is a good candidate. Loosely following Strevens (2008), Levy and Bechtel (2013) claim that a model can explain by describing only the properties of a mechanism that constitute the minimum conditions sufficient to produce the *explanandum*. They show that the explanatory power of some biological models that provide very abstract descriptions of the causal organization underlying a given *explanandum* can be accounted for by this criterion. Nevertheless, merely showing that it can legitimize some abstract mechanistic models does not suffice to argue that it legitimizes mechanistic computation. First, we must prove that the criterion does not motivate the omission of causal organization specifically in computational models. I address this point in section 4. Second, we must show that the criterion legitimizes neurocognitive models abstract enough to be computational. To appreciate the relevance of this second point is important to note that the standard description of vehicles of neural processing is, *in fact*, medium independent (Piccinini and Bahar 2013).

Piccinini and Bahar (2013) proposed an interesting argument to show that neural processing is computational in the generic sense. The idea is that since (i) neural processes are defined over medium independent vehicles and (ii) the processing of medium independent vehicles constitutes

computation in the generic sense, then it follows that (iii) neural processes are computations in the generic sense. Premise (ii) is the result of Piccinini and Bahar's elucidation of generic computation. They only need to argue for premise (i). They point out that current evidence indicates that the primary vehicles manipulated by neural processes are neuronal spikes (action potentials) and that the functionally relevant aspects of neural processes depend on dynamical aspects of spikes such as spike rates and spike timing. Only these dimensions of variation of their vehicles, not any more specific property, are functionally relevant for neural processes. This means that the vehicles of neural processing are medium independent.

Of course, neural processing also involves specific physical properties, such as the electrical processes occurring in the neuron's membrane and the chemical processes that allow for communication between neurons. Neural processing can be described at different levels of abstraction, and a certain level (that involving only spike timing or rate) is medium independent. However, the mere fact that there is a medium independent description of neural processes available, or even the fact that that it is usually employed in cognitive neuroscience, does not settle the question of whether models that make use of this abstract description are *explanatory*. Mechanism is a normative position regarding explanation. The 3M requirement (or even MDB) mentioned above determines that, even if it is found in the neurocognitive literature, a model that does not satisfy this requirement is not explanatory. If we incorporate Levy and Bechtel's criterion as part of our mechanistic approach, we must show that it counts computational neurocognitive models as explanatory. In the following section, I address this point. I argue that Levy and Bechtel's criterion can account for the explanatory power of neurocognitive models that are computational in Piccinini's sense. In particular, I maintain that the criterion legitimizes the abstractions in recent computational models that describe so-called "canonical neural computations" and that these abstractions involve all relevant non-computational features, according to Piccinini's notion. Nevertheless, I claim that the application of the criterion to these models suggests some modifications of Piccinini's notion of neural computation.

3. Minimal conditions and neural computation

As Piccinini (2015), Levy and Bechtel (2013) have also recently argued that mechanism does not imply a commitment with MDB. They consider some abstract models in biology and suggest a criterion that could motivate the relevant abstractions. They focus on a set of models developed by Uri Alon (2007a, 2007b) and colleagues to explain the regulation of gene expression, principally in bacteria and yeast. The models describe the causal organization underlying this behavior in an abstract way, employing a set of tools from graph theory. These models use nodes to represent the components of a given mechanism and edges to represent their operations. These graphs can contain very little information about components and activities. A node usually only specifies some basic response properties of a component regarding other elements (especially the conditions under which it becomes active). Edges typically represent no more than the direction and magnitude of the interaction between two nodes. Thus, graph-based models represent connectivity in different mechanisms in a similar manner even when parts and operations vary with respect to many of their concrete properties. With these tools, Alon models patterns of connections between small numbers of units that have distinctive consequences for the behavior of a biological network. Alon calls these patterns "network motifs."

Graphs are employed also in neuroscience to model different wiring patterns in neural networks. The response properties of a given neuron often depend on the exact wiring pattern of the network it is embedded in. These properties are usually explained by describing the relevant wiring pattern in an abstract manner; representing the different cells as nodes and their inhibitory or excitatory influence on each other as edges. The response of a neuron can be explained, for example, by a graph describing a very simple feedback circuit involving a principal cell and an interneuron. In a feedback inhibitory circuit, increased firing of the principal cell elevates the interneuron's discharge frequency which, in turn, may decrease the principal cell's output, providing a regulatory mechanism similar to that of a thermostat (Jonas & Buzaki 2007). This wiring pattern can be represented in a very schematic way by a graph in which the ending of each edge represents the direction and mode (inhibitory or excitatory) of influence

and the shape of the node represents the kind of cell (principal cell or interneuron) (fig. 1).

More complex firing patterns result when the complexity of the network is increased. One common circuit recruited in different sense modalities (Yantis 2014) is an extension of feedback inhibition. Lateral inhibition occurs when the activation of a principal cell recruits an interneuron, which, in turn, suppresses the activity of surrounding principal cells. This kind of inhibition can result from different specific features of the relevant connections. If principal neurons A and B have a common input and also share a common inhibitory interneuron, a spike train of neuron A can prevent the spiking of neuron B when the input to principal cell A is stronger than the input to principal cell B, when the interneuron-principal cell B synapse is slightly stronger than the interneuron-principal cell A connection or when the input to neuron A arrives slightly earlier than the input to B (Jonas & Buzaki 2007). But it is the abstract wiring pattern (fig. 2) defining lateral inhibition and not these more specific features what accounts for the increased autonomy by competition, or the non-linear "winner-take-all" process that the circuit generates.

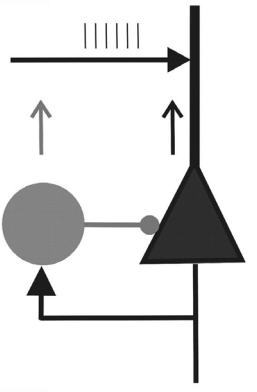

Fig. 1. From Jonas & Buzaki (2007), the graph for an inhibitory feedback circuit involving an interneuron and a principal cell.

One may ask why these graph based models are not less explanatory than more detailed descriptions of the relevant mechanisms. After all, the specific properties (such as the strength of the interneuron-principal cell B synapse)

omitted from the general description of lateral inhibition are explanatorily relevant according to common mechanistic criteria. For example, Craver (2007) provides a criterion to determine when a part of a system S is a component of the mechanism that explains S`s ψ-in: "The hubcaps, mud-flaps, and the windshield are all parts of the automobile, but they are not part of the mechanism that makes it run. They are not relevant parts of that mechanism. Good mechanistic explanatory texts describe all of the relevant components and their interactions, and they include none of the irrelevant components and interactions." (Craver 2007, p. 140)

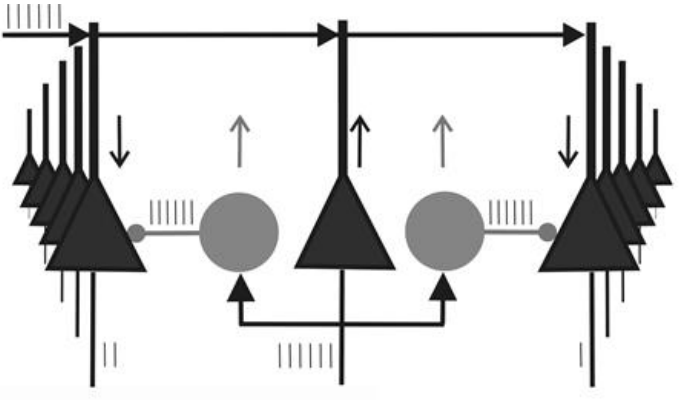

Fig 2. From Jonas & Buzsaki (2007), the schematic representation of lateral inhibition, which allows independence of neural response by suppressing the activity of neighboring neurons ("winner take all").

Craver builds his proposal regarding constitutive explanatory relevance considering the limitations of experimental strategies that neuroscientists employ to determine whether an entity, activity, property or organizational feature is relevant to the behavior of the mechanism as a whole. He considers that constitutive relevance can be determined by the combination of bottom-up and top-down interventions on a mechanism. Bottom-up interventions have limitations that top-down interventions can help to overcome. Bottom-up experiments that Craver mentions are interference and stimulation experiments. In the firsts one intervenes to diminish, disable, or destroy some putative component in a lower-level mechanism and then detects the results of this intervention for the explanandum

phenomenon. The underlying supposition is that if X's φ-ing is a component in S's ψ-ing, then removing X or preventing it from φ-ing should have some effect on S's ability to ψ. (p. 147). In stimulation experiments one intervenes to excite or intensify some component in a mechanism and then detects the effects of that intervention on the explanandum phenomenon. The assumption is that if X's φ-ing is a component in S's ψ-ing, then one should be able to change or produce S's ψ-ing by stimulating X.(p. 149). Although I will not enter into the details of the proposal and its motivation, is important to point out that these experiments are not sufficient to determine whether a component is explanatory relevant because of two limitations. First, sometimes mechanisms compensate for an inhibitory or excitatory intervention and so the intervened component has no effect on the behavior of the system. Second, an intervention can influence the behavior of a system indirectly. These facts make necessary to complement the approach to constitutive relevance with top-down interventions. I will not enter here in the debate on constitutive explanatory relevance. I want only to emphasize that, according to Craver, excitatory and inhibitory bottom-up interventions are relevant (although not sufficient) to determine the explanatory relevance of the component of a mechanism. It is clear then that an inhibitory or excitatory bottom-up intervention on, for example, the strength of an interneuron-principal cell synapse could modify the behavior of a circuit that exhibits lateral inhibition. Therefore, Craver's criterion implies that this information must be included in the model of such a circuit.

However, Levy and Bechtel (2013) propose an abstraction criterion that could motivate or legitimize the abstractions involved in these kinds of graph based mechanistic models. They consider that some models aim to track the features of a system that make a difference to the behavior being explained (p. 256). These models can explain by describing the *minimum conditions that constitute the organizational schema sufficient to produce a given behavior:* "Altering the details of the components (as long as they meet the minimum conditions for fulfilling the role in the organizational schema) does not change the behavior, whereas altering the organization (changing what is connected to what) does" (p. 253). Their main idea is that an explanatory model can omit all the properties of a given mechanism that

can be changed or replaced without modifying the behavior of the system as a whole. In what follows, I will call this criterion "LB."

We saw that lateral inhibition of a principal neuron B by a principal neuron A can be produced by different properties of the circuit they constitute. Although the modification of the strength of the interneuron-principal cell B synapse can prevent lateral inhibition from occurring (and therefore is a feature that has constitutive relevance), this behavior can also be produced by manipulating the relative time of A's and B's inputs. Given that the strength of the connections can be altered without modifying the behavior of the system, this information can be omitted from the model, according to LB.

In what follows I will determine whether LB can legitimize some abstract neurocognitive models that describe computational features of neural processing. In particular, I will focus on the case of the so-called "canonical neural computations" (CNCs). Chirimuuta (2014) argues that the abstractions involved in these models are motivated by a non-mechanistic framework, by the rules of a kind of optimality explanation: efficient coding explanation. I show that without appealing to the framework of optimality or efficient coding explanations, but only to LB, we can account for the explanatory power of CNC models. This means that there are mechanistic computational neurocognitive models.

Normalization is perhaps one of the most studied CNCs. The normalization model is a model that mathematically explains how simple cells in the primary visual cortex respond to specific stimuli in specific orientations (Heeger 1992). One important phenomenon that the model is intended to explain is cross-orientation suppression (COS) (Bonds 1989). COS is the phenomenon that when a non-preferred stimulus of a simple cell in V1 (e.g. a vertical bar) is presented at the same time as the preferred stimulus (e.g., an horizontal bar), the response of the cell is smaller than its response to the preferred stimulus alone. This phenomenon implies that simple cell response is non-linear and therefore cannot be accounted for by the original model proposed by Hubel and Wiesel (1962). The basic idea of Heeger's model 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 an

equation that constitutes the normalization model:

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

Where Ēi 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 is the ΣE is the sum of responses of all simple cells in the local population. The normalizer term ΣE in the denominator is what explains phenomena such as COS. Carandini and Heeger (2012) present normalization as a CNC. These are defined as standard computational modules that apply the same operations in a variety of contexts. 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.

I mentioned earlier that LB motivates the omission of properties or components of a mechanism that can be changed or replaced without modifying the behavior of the system as a whole. Therefore, we can determine that LB legitimize the normalization model if there is a kind of behavior that, in different systems, results from normalization and if normalization is implemented in these different systems by circuits that have no relevant similarities. This would imply that all features other than normalization can be modified without changing the behavior of the system, i.e., that the normalization model captures the minimum conditions sufficient to produce such behavior. As it happens, systems that exhibit normalization fulfill these conditions. First, normalization is implemented by completely different mechanisms in different systems. For example, shunting inhibition and synaptic depression are completely different mechanisms that implement normalization in different brain regions. Synaptic depression is a form of short term plasticity, i.e., the phenomenon that synaptic efficacy changes over time in a way that reflects the history of presynaptic activity. Synaptic depression is caused by depletion of neurotransmitters consumed during the synaptic signaling process at the axon terminal of a pre-synaptic neuron (Tsodycs & Wu 2013). On the other side, shunting inhibition is generated by inhibitory synapses located

close to the soma of a post-synaptic neuron. These synapses increase post-synaptic membrane conductance and therefore locally reduce the input resistance (and thus spiking) (Silver 2010). These different neural mechanisms do not share any relevant property besides the fact that they can produce an inhibitory divisive effect on a post-synaptic neuron. Second, despite these differences, mechanisms that implement normalization can serve the same function, exhibit the same behavior. Normalization helps to maintain the specific calibration of simple cells regarding a small range of stimulus orientations, independently of stimulus contrast (Heeger 1992). Normalization maintains fixed stimulus selectivity.

These two points imply that in the same way as, for example, lateral inhibition regarding a winner-take-all operation, *normalization constitutes the set of minimum conditions that a system has to fulfill to maintain stimulus selectivity.* The information omitted (e.g., the description of the shunting inhibition mechanism) is about features that can be changed without modifying the behavior of the circuit. If this is so, the normalization model is an explanatory abstract model according to LB.

What is most important for our present purposes is that the abstractions in the normalization model that can be legitimized by LB are enough to make the description medium independent, i.e., computational. The information that LB can determine as irrelevant for the *explanandum* is precisely the information about the specific physical processes that underlie normalization. The mechanism of shunting inhibition, for example, is a mechanism whose activities constitute a specific electrical process, namely, a short-circuit generated by an inhibitory neuron in the post-synaptic membrane. Shunting inhibitory synapses are often located close to the soma, where their conductance can have a large effect on somatic input resistance (and thus spiking) because of the proximity to the spike initiation zone. As mentioned above, the increase in membrane conductance that these synapses introduce short-circuits excitatory synaptic currents by locally reducing the input resistance. These shunting inhibitory conductances scale down post-synaptic excitatory potentials in a multiplicative manner, in accordance with Ohm's law. Thus, LB legitimizes the abstraction of all specific physical processes and principles underlying, for example, stimulus selectivity, and therefore makes the model that explains it medium

independent.

An important consequence of using LB as a criterion to legitimize computational neurocognitive models is that it can help us to clarify the notion of neural computation itself. As we have seen, Piccinini and Bahar (2013) claim that neural processes are computational because their vehicles are medium independent. They support this claim by stating that spike time and rate are the vehicles of neural processing and that these are medium independent. But the normalization model describes neural computational processes without specifying a coding regime (without specifying 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, & McCormick 2003). Therefore, coding regime is relevant to characterizing neural circuits and not computations.

It is important to emphasize that this does not mean that there is a single computational level in neural mechanisms[4]. Medium independent descriptions are appropriate for different levels of a neural mechanism. Many cortical areas and other large neural systems that perform computations have components that also compute. The more limited computations that columns and nuclei perform are component processes of the computations performed by their containing systems. There are different computational levels before we reach a purely biophysical level (Boone and Piccinini 2015). Coding regime independence may constitute

[4] I thank Piccinini (personal communication) for pointing out that, in order to work, my argument must not have this implication.

one or more computational levels. However, neural code is not part of any of these computational levels if it is only relevant (i.e., it is a difference maker) for processes that are not medium independent (such as shunting inhibition). To show that neural code belongs to a computational level we should determine its relevance for a task that does not involve specific physical properties or principles.

To conclude this section, it is worth noting that omitting information about circuits underlying neural computation makes that information available for *other* explanatory uses which are different from explaining certain informational tasks. For example, this information can be incorporated by an inferior-level model to explain neural computation itself. This is exactly the use of neural circuits we find in the literature. For example, there are specific features of shunting inhibition that explain why the inhibition is specifically *divisive*. Classical theoretical work (e.g., Blomfield 1974 and Vu & Krasne 1992) suggests that the arithmetic operations resulting from shunting inhibition depend on the size and location of the conductance: inhibition may have a divisive effect on the EPSP if the conductance is large and located close to the soma, but may have a subtractive effect if the conductance is small and spatially distributed (fig. 2).

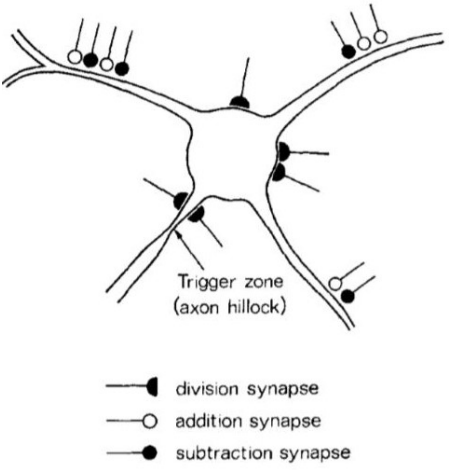

Fig. 3. From Blomfield (1974), shunting inhibition may have a divisive effect on the EPSP if the conductance is large and located close to the soma.

At least regarding the case of neural arithmetical operations, the features omitted from the computational explanation of some informational tasks can be recycled by the model that explains neural computation itself. Mechanisms such as shunting inhibition are part of a hierarchy in which circuits explain computations that explain informational tasks. This is relevant because *integration* is a central feature of mechanistic explanation. Mechanism is opposed to the view that abstract computational explanations of cognitive phenomena are isolated from low-level descriptions of underlying neural processes. The multilevel approach (see Boone & Piccinini 2015) affirms that neurocognitive explanation is constituted by an integrated hierarchy of models that address different mechanistic levels. At least for the model we are considering, LB is compatible with this characteristic. The information whose omission it motivates can be recycled by connected lower-level explanations.

4. Mathematical descriptions and mechanistic computation

In the previous section, I argued that LB legitimizes some computational neurocognitive models. However, this does not imply that these models are mechanistic. The criterion states that an explanatory model describes the minimum conditions sufficient to exhibit a given behavior. In the case of lateral inhibition (as in the case of the biological models considered by Levy and Bechtel), these conditions are the abstract pattern of *causal* connections described by a graph. But perhaps this is not necessarily so. The minimum conditions for some behaviors could be given by a pattern of, for example, purely mathematical relations. If this is the case of the normalization model considered above, then we cannot affirm that LB legitimizes a *mechanistic* computational model. In this section, I show that LB motivates a mechanistic interpretation of normalization.

Chirimuuta (2014) considers that a distinction must be made between models that describe CNCs and models that describe neural mechanisms. She does not deny that mechanistic models can be abstract. Chirimuuta considers that there is a class of abstract models, which she calls "A-minimal models," that can provide mechanistic explanations. But she insists that models that describe CNCs are not abstract in this sense and that they

are non-mechanistic. Chirimuuta points out that the distinction between computation and mechanism is endorsed by Carandini and Heeger themselves (p. 141). This distinction is in fact frequently employed in the neurocognitive literature, but I believe that it could be motivated also by the idea that mechanistic explanation requires specification of causal connectivity. It seems that the normalization model, for example, is a purely mathematical and not a causal description of the behavior responsible for stimulus selectivity. Normalization is apparently represented as an equation in which the normalized response Ēi of a simple cell is equal to the input from the LGN (or the non-normalized response Ei) divided by the sum of responses ΣE of all simple cells in the local population. Furthermore, Carandini and Heeger show that normalization can be implemented by different *abstract* mechanisms. They describe different abstract wiring patterns that can implement the equation.

As Carandini and Heeger point out (p. 58), it is well known to electrical engineers that gain control can be implemented using either a feedforward or a feedback system. A feedback circuit has been traditionally proposed for primary visual cortex, where signals in the denominator of the normalization equation have been thought to originate from lateral feedback within V1, or from feedback from higher visual areas (fig. 4 b). However, normalization could also be implemented by a feedforward network that taps the non-normalized signals before they have been subjected to normalization. Such an arrangement has been proposed for the visual system of the housefly, for the olfactory system of the fruitfly and for some aspects of normalization in the mammalian visual cortex (fig. 4 a).

---

[5] The question of whether the normalization model includes all the information needed for a fully explanatory mechanistic model (i.e., whether the abstractions involved diminish explanatory power) is different from the question of whether the normalization model can be part of non-mechanistic explanations (such as an optimality or efficient coding explanation). I agree with Chirimuuta (2014) that the normalization model can be employed by non-mechanistic explanations of *why* a system presents a given behavior or trait. What I claim (against Chirimuuta) is that the model includes enough information to be considered fully explanatory (i.e., that it is not merely a mechanism sketch). For example, when Heeger 1992 uses the model to explain- among other things – *how* SOC is performed, this can be taken to be a fully explanatory mechanistic model.

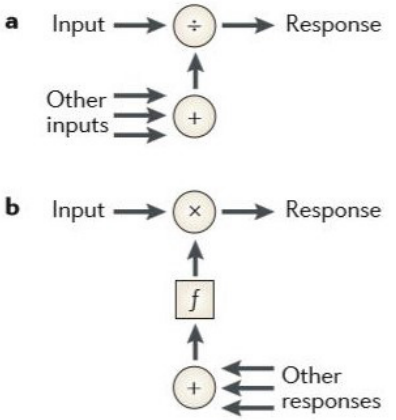

Fig 4. From Carandini & Heeger (2012), feedback and feedforward connectivity patterns can implement the neural gain control that constitutes normalization.

If LB legitimizes the normalization model and this model omits information about the relevant connectivity pattern, i.e., about the abstract causal organization that can be described by a graph, then we cannot say that LB legitimizes a mechanistic computational model. Nevertheless, in what follows, I will show that if normalization is interpreted in purely mathematical terms, then it is not legitimized by LB or, alternatively, that it is legitimized only if we endorse a mechanistic interpretation of the model. To demonstrate this point, I will briefly characterize the arithmetic operations performed by neurons and consider different ways of describing them[5].

Given three responses N_1, N_2, and N_3 from three different neurons or neural populations, these perform an arithmetic operation if N_1 is a response driven by N_2 and if the input-output relation between N_1 and N_2 is modulated by N_3 (Silver 2010). An addition occurs, for example, when N_3 modulates the relation between N_1 and N_2 in an additive way, i.e., when N_3 *excites* (causes an increase in the value of) N_1 in such a way that the value of N_1 is equal to the value of the driving input N_2 *plus* the value of the modulating input N_3.

If the description of the relation between the responses of three neurons is

given by the equation
(1) $N_n + N_m = N_l$
And (1) is understood in a purely mathematical manner, i.e., abstracting away from the fact that N_l is the input of a process that has N_n and N_m as inputs; representing only the mathematical relations between those three variables, then N_1, N_2, and N_3 satisfy (1) but also satisfy the equation:
(2) $N_n = N_l - N_m$

If the description is purely mathematical, then we can make all the transformations that this kind of description allows. Now, let us suppose that we have three neural responses N_4, N_5, and N_6 that do not form a excitatory circuit, as N_1, N_2, and N_3, but an inhibitory one, in which N_6 subtractively modulates the input-output relation between N_4 and N_5, i.e., in which N_6 inhibits (causes a decrease in the value of) N_4 in such a way that the value of N_4 is equal to the value of the driving input N_5 minus the value of the modulating input N_6. As (1) and (2) represent only mathematical relations, responses N_4, N_5, and N_6 (in the same way as N_1, N_2, and N_3) satisfy these two equations. This implies that a purely mathematical description of a neural arithmetic operation is insufficient to determine whether the circuit performs an inhibitory or an excitatory arithmetical operation.

This purely mathematical description can be opposed to description that represents a causal processes that implements a mathematical function. In this description, the terms on one side of the "equation" are the *inputs* of the process, and the result on the other side is the *output*. Under this interpretation, "equation" (1) can represent the process of additive modulation that has N_l as output and N_n and N_m as inputs. The same process cannot be represented by certain equations equivalent to the mathematical interpretation of (1). For example, the process that implements (1) does not implement (2). This process has N_l and not N_n as an output, and the output is the sum and not the difference of the inputs. Under this interpretation, $N_n + N_m = N_l$ can represent the circuit composed of N_1, N_2, and N_3, but not the one composed of N_4, N_5, and N_6. In the same way, $N_n = N_l - N_m$ can represent the circuit composed of N_4, N_5 and N_6 but not that composed of N_1, N_2, and N_3.

For the same reason, if a normalization equation represents a mathematical and not a causal structure, then it can be satisfied by a

circuit in which the response of a neuron is not divisively inhibited but multiplicatively excited. The variables in this circuit maintain mathematical relations equivalent to those between the variables in the normalization equation. However, the circuits that present this causal organization cannot explain some inhibitory phenomena, such as COS, that normalization is supposed to explain. As we have seen, COS is a phenomenon which occurs when a response is *decreased* when the preferred stimulus is presented with a non-preferred stimulus. This implies that the modulatory input must be inhibitory. Therefore, if normalization constitutes the minimum conditions sufficient to produce COS, then it cannot describe merely a mathematical organization, but causal relations and their quantitative properties. This is why LB motivates a mechanistic interpretation of the modeling of neural arithmetic operations that define different CNCs.

If this is correct, the models that describe canonical neural computations can be considered both mechanistic and computational descriptions of neural processes. Although this argument concerns a limited set of neurocognitive models, its relevance can be enhanced if we emphasize the fact that the small set of neural operations that constitute canonical computations has a widespread relevance for cognition. Carandini and Heeger (2012) present normalization as a *canonical* neural computation. These computations are defined as standard computational modules that perform the same operations in a wide variety of contexts (p. 51). Normalization was proposed in the early 1990s to explain non-linear properties of neurons in the primary visual cortex. However, evidence that has been accumulated since then suggests that normalization plays a part in a wide variety of modalities, brain regions, and species. The normalization model has been successfully applied to a very wide variety of neural systems: 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. This means that providing an argument for a mechanistic characterization of canonical computations implies an important step towards a general mechanistic framework for neural computation.

5. Conclusion

I have argued that we must provide an adequate abstraction criterion if we wish to claim that (at least some) neurocognitive models are, at the same time, computational and mechanistic. I have shown that the LB criterion is a promising candidate in this respect. First, I argued that the abstractions it motivates in some neurocognitive models are sufficient for medium independence. This point is important because although medium independent descriptions are employed in neurocognitive models, as a normative approach, mechanism must determine whether these abstractions are well motivated. Furthermore, a consequence of the application of LB to the modeling of neural computation is that this is not only medium independent but also coding regime independent. Second, I argued that LB can also motivate a mechanistic interpretation of the considered models. Although there is a common distinction between neural computation and neural mechanisms, I maintained that LB legitimizes the computational normalization model only if we consider that it includes all the information necessary for mechanistic explanation, and that this interpretation is in consonance with the current neurocognitive interpretation.

It is important to emphasize that although the claims defended here intend to blur the distinction between computational and mechanistic explanations (and therefore to expand the scope of the mechanistic approach) they are not in conflict with a pluralistic perspective about neurocognitive explanation. Furthermore, they can constitute a contribution to pluralism as they can help to draw more accurate distinctions between different kinds of explanations. Specifically, they can show that although there is a set of non-mechanistic neurocognitive explanations (such as the efficient coding explanations considered by Chirimuuta 2014), not all computational explanations are part of this set.

Acknowledgements

I thank Gualtiero Piccinini for his comments and inspiring work on neural computation. I am also grateful to the referees from the Journal of

Cognitive Science for very insightful comments. Finally, I want to thank Liza Skidelsky, Sergio Barbieris, Sabrina Haimovici, Mariela Destéfano, Fernanda Velazquez, Nicolás Serrano, Magali La Rocca and Rodrigo Gonzáles Wilkens and Cristian Stábilefor exhaustive discussions on many versions of this paper.

References

Alon, Uri. (2007a) *An Introduction to Systems Biology: Design Principles of Biological Circuits.* Boca Raton, FL: Chapman & Hall

Alon, Uri. (2007b) Network Motifs: Theory and Experimental Approaches. *Nature Reviews Genetics* 8:450–61.

Amos, S. W. y James, M. (2000) *Principles of Transistor Circuits: Introduction to the Design of Amplifiers, Receivers, and Digital Circuits*. Oxford, Elsevier Science & Technology

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

Blomfield, S. (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, W. y Piccinini, G. (2015) The cognitive Neuroscience Revolution. *Synthese:* 1-26. Published online, DOI: 10.1007/s11229-015-0783-4.

Boone, W. y Piccinini, G. (forthcoming in *Philosophy of Science*) Mechanistic Abstraction.

Carandini, M., y Heeger, D. J. (2012) Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13, 51–62

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

Craver, C. F. (2007) *Explaining the brain: Mechanisms and the mosaic unity of neuroscience.* Oxford: Oxford University Press.

Haimovici, S. (2013) A Problem for the Mechanistic Account of Computation. *Journal of Cognitive Science*, vol. 14: 151-18

Garson, J. (2003) The Introduction of Information into Neurobiology. *Philosophy of Science* 70 (5):926-936

Heeger, D. J. (1992) Normalization of cell responses in the cat striate cortex. *Visual*

Neuroscience, 9, 181–197

Hubel, D. H., and Wiesel, T. N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106–154

Jonas, P. and Buzsaki, G. (2007) Neural inhibition. *Scholarpedia*, 2(9):3286.

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

Lange, M. (2013) What makes a scientific explanation distinctively mathematical?.

British Journal for the Philosophy of Science, 64, pp. 485-511.

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

Piccinini, G. and Bahar, S. (2013) Neural Computation and the Computational Theory of Cognition. *Cognitive Science* 37 (3), 453-488.

Piccinini, G. (2015) *Physical Computation, a Mechanistic Account*. Oxford, Oxford University Press.

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

Silver, R., A., (2010) Neuronal arithmetic. *Nature Reviews Neuroscience* 11: 474-489.

Tsodyks, M. and Wu, S. (2013) Short-term synaptic plasticity. *Scholarpedia*, 8(10):3153.

Vu, E. T. & Krasne, F. B. (1992) Evidence for a computational distinction between proximal and distal neuronal inhibition. *Science* 255, 1710–1712

Yantis, S. (2014) *Sensation and Perception*. New York, NY: Worth Publishers