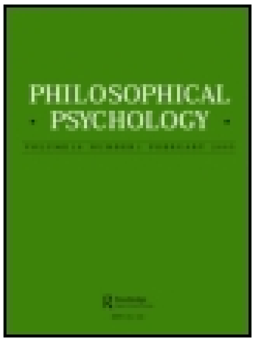

Routledge
Taylor & Francis Group

Philosophical Psychology

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: http://www.tandfonline.com/loi/cphp20

A mechanistic perspective on canonical neural computation

Abel Wajnerman Paz

To cite this article: Abel Wajnerman Paz (2017): A mechanistic perspective on canonical neural computation, Philosophical Psychology, DOI: 10.1080/09515089.2016.1271117

To link to this article: http://dx.doi.org/10.1080/09515089.2016.1271117

Published online: 01 Jan 2017.

PHILOSOPHICAL PSYCHOLOGY, 2017
http://dx.doi.org/10.1080/09515089.2016.1271117

Routledge
Taylor & Francis Group

A mechanistic perspective on canonical neural computation

Abel Wajnerman Paz

University of Buenos Aires, National Scientific and Technical Research Council, Buenos Aires, Argentina

ABSTRACT

Although it has been argued that mechanistic explanation is compatible with abstraction (i.e., that there are abstract mechanistic models), there are still doubts about whether mechanism can account for the explanatory power of significant abstract models in computational neuroscience. Chirimuuta has recently claimed that models describing canonical neural computations (CNCs) must be evaluated using a non-mechanistic framework. I defend two claims regarding these models. First, I argue that their prevailing neurocognitive interpretation is mechanistic. Additionally, a criterion recently proposed by Levy and Bechtel to legitimize mechanistic abstract models, and also a criterion proposed by Chirimuuta herself aimed to distinguish between causal and non-causal explanation, can be employed to show why these models are explanatory only under this interpretation (as opposed to a purely mathematical or non-causal interpretation). Second, I argue that mechanism is able to account for the special epistemic achievement implied by CNC models. Canonical neural components contribute to an *integrated* understanding of different cognitive functions. They make it possible for us to explain these functions by describing different mechanisms constituted by *common* basic components arranged in different ways.

ARTICLE HISTORY
Received 17 February 2016
Accepted 22 November 2016

KEYWORDS
Abstraction; causal explanation; cognitive neuroscience; integration; mechanism

1. Introduction

Mechanism is an influential view about scientific explanation whose main thesis is that a model explains a given phenomenon only if it accurately describes the mechanism underlying it. This thesis has been interpreted by both advocates (e.g., Craver & Kaplan, 2011; Kaplan, 2011; Kaplan & Craver, 2011) and critics (e.g., Barberis, 2013; Haimovici, 2013) of mechanism as implying that the explanatory power of a model is increased as we include more relevant information about the mechanism underlying its target phenomenon. Against this claim,

CONTACT Abel Wajnerman Paz abelwajnerman@gmail.com

mechanists (e.g., Boone & Piccinini, 2016b; Levy & Bechtel, 2013) have argued that mechanism is compatible with abstraction or, in other words, that abstract models can be considered fully explanatory. However, there are still doubts about whether mechanism can account for the explanatory power of relevant abstract models in computational neuroscience. Chirimuuta (2014) has recently claimed that we need a non-mechanistic framework to understand how models describing canonical neural computations (CNCs) explain. CNCs are computational modules that are implemented recurrently in various neural systems to perform different informational tasks. Given the wide impact that their canonical character implies for our understanding of neural processing, these models deserve a thorough evaluation.

In this paper, I defend two claims. First, I argue that the prevailing interpretation of these models in computational neuroscience is mechanistic and that there are good reasons to conclude that they are explanatory only under this interpretation (as opposed to a purely mathematical one). Second, I claim that mechanism is able to account for the special contribution of CNC models to the study of neural processing. The discovery of CNCs makes it possible for us to develop an integrated understanding of many different cognitive functions. I will argue that this integration can be accounted for by the componential character of mechanistic explanation.

In section 2, I present the mechanistic approach to explanation and its presumed implication regarding the relation between explanatory power and mechanistic detail. I describe the criterion proposed by Levy and Bechtel (2013) as a way to elude this implication within a mechanistic framework. Finally, I consider an argument made by Chirimuuta (2014) aimed at showing that even if we accept mechanistic abstraction, mechanism is not able to account for the explanatory power of abstract models that describe CNCs. I offer a more elaborate version of this argument, employing Chirimuuta's more recent characterization of non-causal explanation. Chirimuuta (in press) takes purely mathematical models considered by Lange (2013) as paradigm cases of non-causal explanation. In section 3, I argue that the prevailing interpretation of CNC models is not purely mathematical in this sense but involves the description of causal relations between components of a mechanism. Also, I claim that these models can be considered explanatory only under this causal interpretation. Counterfactual differences in CNCs are *difference makers* for the relevant explananda only because CNCs are constituted by causal and not merely mathematical relations.

In section 4, I argue that mechanism can account for the special epistemic achievement implied by CNC models. It has been argued that one significant virtue of mechanism is that it can account for the increasing integration between different explanations produced by the development of cognitive neuroscience (Boone & Piccinini, 2016a). I argue that CNC models can be embedded in a hierarchy of mechanistic explanations of the kind considered by Boone and Piccinini

(2016a) and, more importantly, that they can also contribute to a horizontal integration of models at the same level or degree of abstraction. The discovery of canonical components provides a language to describe different mechanisms and explain different tasks in a unified manner, employing (the description of) a common set of basic operations. This kind of integration, which is taken by proponents of CNCs to constitute their special epistemic value, can be accounted for by the componential character of the mechanistic explanations in which CNCs are involved.

2. Mechanism, abstraction, and causal organization

To understand the relation between CNCs and the mechanistic framework, it is crucial to first 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 & 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).

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:

> **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, 2011, p. 347)

Although I will not enter into the debate about the characterization of the notion of model, a few terminological clarifications are in order. I will follow the authors relevant for the present discussion (e.g., Boone & Piccinini, 2016a; Chirimuuta, 2014; Chirimuuta, in press; Kaplan, 2011; Levy & Bechtel, 2013) in using the term "description" to refer to models. I will assume that (at least some) models are descriptions of a relevant mechanism. I interpret the 3M requirement as implying the idea that some explanatory models can be descriptions constituted by a set of symbols and relations between them whose parts can be mapped onto properties of a target mechanism. I will not defend this "descriptivist" view of models here. I will take it as a common background assumption shared both by advocates and critics of mechanism in the present discussion.[1]

A further notion that is relevant for what follows is that of abstraction. I will follow Levy and Bechtel (2013) and Piccinini (2015) in the idea that abstraction is mere information omission. To relate this idea to the 3M requirement, I will not use “information” in some usual technical sense (as in Shannon’s information theory) but I will identify the information a model carries with the set of mapping relations between model parts and target system properties. When I say that a given model is abstract regarding a given property *P* (that it omits information about *P*), I mean that there is no part of the model that maps onto an instance of *P* in the target system or mechanism.[2]

Kaplan (2011) considers mechanism to have a commitment in addition to 3M. 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, for example, including additional variables to represent (or that map onto) additional components of the mechanism. Chirimuuta (2014) calls this requirement *the more details the better* (MDB). MDB implies that models that involve some kind of abstraction, that omit some information about the target mechanism (e.g., models that describe only high level properties), are less explanatory than more detailed descriptions. This is problematic because there are many models that seem to be fully explanatory despite the fact that they constitute very selective depictions of the mechanisms underlying relevant phenomena.

For instance, Levy and Bechtel (2013) have recently considered a set of abstract models developed by 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. On a graph, the components are represented as nodes and their operations as edges. In many cases, a node only indicates that a certain element in the system exists and that it has some basic response properties 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 and lack many other pieces of information (e.g., whether the interaction is mechanical, chemical, or electrical). With these tools, Alon models patterns of connections among a small numbers of units that have distinctive consequences for the behavior of a biological network. He calls these patterns “network motifs.”

Alon proposes, for example, a graph to represent the mechanism underlying regulation of arabinose in bacteria. This mechanism is represented by three nodes and three edges. An initial transcription factor X regulates a second transcription factor Y, and both of them regulate an operon Z that synthesizes the enzymes employed in arabinose metabolism. As Z requires both X and Y to be activated, the relation between Z, on the one side, and X and Y, on the other, is represented by a Boolean AND gate (Figure 1). cAMP (Sx) signals the absence of glucose (and therefore the need to use arabinose), but this signal is noisy. Pulses of cAMP are briefly triggered by the bacterium transitions between different growth conditions.

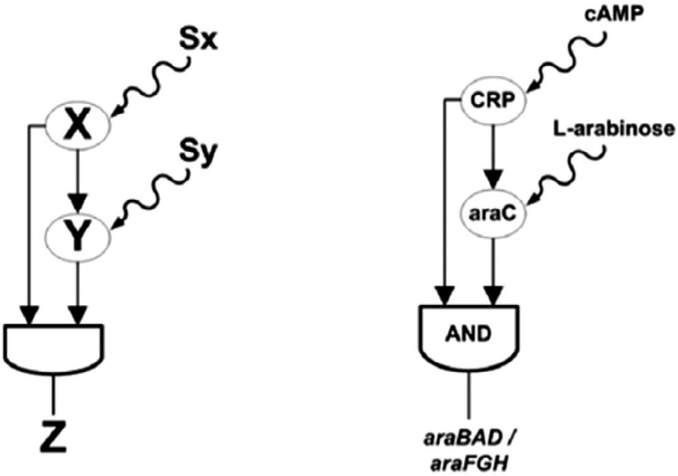

Figure 1. From Mangan and colleagues (2003). The graph represents the motif responsible for arabinose regulation.

Only when the cAMP signal is longer than the typical spurious pulse it can trigger not only the transcription factor CRP (X) but also AraC (Y) and therefore cause the arabinose operon (Z) to begin synthesis of key enzymes (Mangan, Zaslaver, & Alon, 2003). As predicted by Alon, the motif is useful in this system as a signal persistence detector.

Levy and Bechtel (2013) make explicit the criterion behind the highly selective depiction of the underlying mechanism that constitutes Alon's model. Based in a general way on Strevens's (2008) approach to abstract causal explanations, they argue that the model aims to track those features of the system that *make a difference* to the behavior being explained, namely, persistence detection. They claim that a model 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). We can say that an abstract model is explanatory if it omits only information about features that can be changed or replaced without modifying the behavior of the system[3]. In what follows, I will call this criterion LB.

Recently, Chirimuuta (2014) has argued that even if we accept that mechanism and abstraction are compatible, there are relevant abstract models in computational neuroscience whose explanatory power cannot be accounted for by this approach. One of the models she evaluates is the normalization model proposed by Heeger (1992). This is 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 nonlinear. 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 (e.g., a 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). The basic idea of Heeger's model is that each simple cell has a linear excitatory input from the *lateral geniculate nucleus* (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:

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

Where $\bar{E}_i$ describes the normalized response of a simple cell, t stands for time, σ^2 characterizes a parameter that governs the contrast at which the neuron is saturated, and ΣE describes the sum of responses of all simple cells in the local population. The normalizer term ΣE in the denominator can explain phenomena such as COS. Carandini and Heeger (2012) present normalization as a *canonical neural computation* (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.

Chirimuuta claims that the explanatory power of the models that describe CNCs cannot be accounted for within mechanism. On the contrary, these models can be considered explanatory only if they participate in efficient coding explanations that cannot be evaluated by mechanistic norms. These are a kind of the optimality explanations often found in biology, which explain a phenomenon by showing that it constitutes the optimal strategy to solve a given problem. This strategy can usually be specified by a purely mathematical description without any reference to its bio-physic implementation (Rice, 2015).

Chirimuuta (2014) concedes that there are abstract explanatory mechanistic models such as that considered by Levy and Bechtel. These are what she calls "A-minimal models," which highlight the most relevant causal features or *difference makers* of a mechanism. But even conceding abstract mechanistic explanations, she insists that models that describe CNCs are non-mechanistic. Therefore, the non-mechanistic character of CNCs cannot be attributed merely to their abstract character but must depend on some form of abstraction that is problematic for mechanism. Indeed, Chirimuuta considers that some features of the normalization model imply that (if it were evaluated by the mechanistic norms for explanation) it should be considered a mechanism sketch. This is a kind of model commonly contrasted with mechanistic models. A sketch is a model that describes some components of the mechanism underlying a phenomenon (and some of their relations) but omits details that are not yet known (Machamer, Darden, & Craver, 2000).

Chirimuuta (2014) points out that the normalization model gives a quantitatively accurate prediction of cross-orientation suppression and numerous other phenomena (Heeger, 1992). The model is able to make these predictions by merely describing the suppressive effect, ΣE, which characterizes the underlying inhibitory mechanism in a very schematic way. However, pace Chirimuuta, this is not sufficient to affirm that the model is a sketch. The information omitted from the model is not unknown. Shunting inhibition and synaptic depression are well studied mechanisms that underlie normalization, and we even know the specific way in which they contribute to its implementation (see section 4). The abstract character of the normalization model is not due to an imprecise knowledge of the relevant mechanism and therefore cannot be considered a mechanism sketch.

Chirimuuta considers a further form of abstraction that is problematic for mechanism. Although she does not address the normalization model, Chirimuuta (in press) claims that neurocognitive models providing efficient coding explanations cannot be considered mechanistic because they are not causal. Even mechanists that defend abstract models consider that an explanatory model cannot omit all information about causal properties of a relevant mechanism. Levy and Bechtel (2013) affirm that mechanistic explanations address *organized* systems. A system is organized with respect to a given behavior of a phenomenon if different components of the system make a different causal contribution to the behavior and the component's differential contributions are integrated (that is, each component interacts in particular ways with a subset of the other components). An explanation is mechanistic only if it describes causal organization, that is, only if it describes the different causal contributions of the mechanism underlying a phenomenon and how these contributions are integrated.

Chirimuuta (in press) develops some of Woodward's ideas to distinguish between causal and non-causal explanation. She claims that Woodward's approach to explanation, which is accepted by mechanists (e.g., Kaplan, 2011; Kaplan & Craver, 2011), can be generalized beyond causal explanation. The main thesis of this interventionist theory of causal explanation is 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 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, in press, p. 5). These changes are characterized as the result of an intervention. An intervention is "an idealized, unconfounded experimental manipulation of one variable which *causally* affects a second variable only via the causal path running between these two variables" (Woodward, 2013, p. 46). It is relevant to notice the similarity between this requirement and LB. As I mentioned, LB affirms that a model can explain by referring only to the features of a mechanism that are difference makers in this sense, that is, features whose modification cause modifications in the *explanandum*.

As Chirimuuta points out, Woodward himself suggests that a model can address w-questions without referring to causal properties (Woodward, 2003, p. 221). The example mentioned by Woodward is that of the hypothesis that the stability of the planets is counterfactually dependent on the four-dimensional structure of space-time. The w-question of "what if space-time had been six-dimensional?" associated with this hypothesis has no associated causal intervention. There seems to be no possible (idealized or otherwise) intervention that could result in the modification of the structure of space-time. But the hypothesis does provide an answer to this question because it implies that if things had been different, then planetary orbits would indeed be less stable. The claim that the description of a CNC is not mechanistic can then be interpreted as the claim that these descriptions do not include causal information to address relevant w-questions.

Although Chirimuuta (2014) does not claim that CNC models provide non-causal explanations, she considers that their explanatory power comes from their ability to address w-questions. More importantly, she gives us an idea of what the relevant w-question for these models should look like. These w-questions concern the set of counterfactual dependences between the input of the system that runs a CNC (e.g., sensory information) and/or the system requirements (e.g., the task for which the information is needed) and the computational properties of the system (e.g., CNCs themselves). The model implies that if the task that the system needs to perform and/or its sensory input were different, then the underlying computations would also be different. Chirimuuta offers the example of a study by Wainwright and colleagues (2002), which shows that the normalization parameters are adjusted by variations in the statistics of recent visual input. In the following section, I will argue that to address w-questions of this kind, the normalization model must be interpreted as providing a causal explanation.

3. Mechanistic abstraction and canonical neural computations

As I mentioned in the previous section, Chirimuuta (in press) points out that the difference makers that can be employed to address relevant w-questions regarding a phenomenon can be non-causal features described by a model. To provide a more accurate characterization of this kind of non-causal model (and compare it with what I take to be the correct interpretation of the normalization model) it is helpful to consider some paradigmatic cases. Chirimuuta takes the distinctively mathematical explanations described by Lange (2013) as typical examples of non-causal explanations.

According to Lange, a mathematical explanation in empirical science is an explanation that depends only on mathematical laws or principles, that is, principles that have a stronger modal force than natural laws. To understand how a

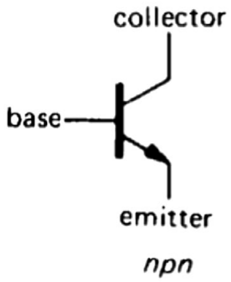

Figure 2. From Horowitz and Winfield (1989). The schematic representation of an NPN transistor.

distinctively mathematical explanation works, let's consider the example of the determination of the mathematical relation between the two current gains β and α of an NPN transistor. In the representation of a transistor circuit, we can distinguish three currents I_c, $I_{b,}$ and I_e that correspond, respectively, to the currents in the collector C, the base B, and the emitter E of the transistor (Figure 2).

The current I_c flows between terminals C and E when the transistor is "on," which happens only when a small current I_b flows into the terminal B. Thus, current I_e in E is given by equation (1):

$$I_c + I_b = I_e \tag{1}$$

The base works as current control input (as a switch). The transistor also possesses a gain β, defined by the equation:

$$\beta = I_c/I_b \tag{2}$$

And another gain α, defined by the equation:

$$\alpha = I_c/I_e \tag{3}$$

From these three equations we can infer the mathematical relation between gains β and α (and thus calculate one from the other) in the following way:

$$I_b = I_e - I_c(\text{from } 1) \tag{4}$$

$$I_b = I_e - (I_e\alpha)(\text{from } 3) \tag{5}$$

$$I_b = I_e(1 - \alpha)(\text{from } 5) \tag{6}$$

$$\beta = I_c/I_e(1 - \alpha)(\text{from 2 and 6}) \tag{7}$$

And finally:

$$\beta = \alpha/1 - \alpha(\text{from 3 and 7}) \tag{7}$$

The relation between gains β and α is mathematically deduced from the way we define them using the transistor currents plus the mathematical relation (1) that the currents maintain.[4] Each step of the derivation is mathematically valid, since it is warranted by some arithmetic principle or axiom. This means that the implication derived is warranted by principles that have a modal force greater than that of a natural law. I agree with Chirimuuta (in press) that the epistemic value of these explanations can be accounted for by Woodward's requirement of addressing w-questions. I also agree that (at least some of) these w-questions do not involve situations that result from causal interventions. We could ask how a transistor would behave if the mathematical principle that leads from step 5 to 6 in the derivation above were false. If this were the case, the relation between the two gains would be different. But this difference could not be attributed to any causal intervention on a component of the mechanism that constitutes the transistor. The relevant mathematical principle is no something on which one could causally intervene.

I consider that CNC models are not mathematical models in this sense. They do not explain by describing purely mathematical relations, but rather causal relations mathematically characterized. To begin, this is their common neurocognitive interpretation. Neural arithmetic is defined by a set of operations in which a modulatory input modifies the input–output relation between two neurons or neural populations in a given direction (Silver, 2010). Given three responses N_1, N_2, and N_3 from three different neurons or neural populations, these responses perform an arithmetic operation only 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. An addition occurs, for instance, when N_3 modulates the relation between N_1 and N_2 in an additive way, that is, 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. Although descriptions of neural arithmetic operations often have the form of mathematical equations, the arithmetic symbols in these descriptions refer to (or are mapped onto) mathematically characterized causal relations between neural responses. The addition symbol "+" does not refer to a mathematical relation that different variables of the target system have, but to a causal relation of additive (as opposed to multiplicative) excitation of the activity of one component by another. This causal interpretation can be seen in pioneering theoretical work on neural arithmetic. For instance, Blomfield (1974) characterizes neural arithmetic as operations that a (modulatory) neuron performs on its synaptic input. This interpretation is also explicit in more recent development on canonical computation. Carandini and Heeger's (2012) review of the different applications of the normalization model shows that in different instances of the equation the denominator refers to an *inhibitory* input, which modulates the output response. In contrast, when a multiplication symbol appears in a CNC equation, it does not refer to a relation that can be equivalent to a divisive relation (as in a purely mathematical description). For instance, in Reynolds and Heeger's (2009) model of attention, multiplication is an *excitatory* response of an attention field that modulates the output firing rate (see section 4).

This causal interpretation of CNCs is not only the prevailing one but also the one that can be said to have explanatory power. This claim can be motivated by the idea underlying LB and Woodward's requisite that an explanatory model must describe the relevant difference makers. In what follows, I will argue that the normalization model does pick up difference makers and then show that these must be interpreted as causal and not purely mathematical properties. As we have seen, the normalization model is employed to explain the phenomenon of *cross orientation suppression* (COS), among others. There are reasons to think that the model excludes information about features that are not difference makers regarding this phenomenon and includes information about the properties that are.

First, the non-computational properties underlying normalization are not difference makers for COS. Mechanisms underlying normalization are not the same in different systems and species. For instance, synaptic suppression and shunting inhibition are different mechanisms that implement normalization in different brain areas. Normalization can produce the nonlinear response of a neural population required by COS even if it is implemented by different circuits or neural mechanisms. Second, if divisive normalization did not affect the response of simple cells in V1, then these would not have the nonlinear properties required to produce COS. As I mentioned, these properties are not predicted by the model proposed by Hubel and Wiesel, which does not include divisive normalization. This means that the model does pick up difference makers regarding COS.

We can interpret the model as mapping onto difference makers only if we take these to be causal, and not merely mathematical, properties. If we interpret normalization as a purely mathematical description, then it follows that it does not refer to difference makers relevant to addressing the w-questions regarding its *explanandum*. COS is an inhibitory phenomenon, in other words, one that occurs when a given input *decreases* a given neural response. This decrease in the value of neural activity is explained by normalization because the model describes an inhibitory circuit. We have seen that although the normalization model has the form of a mathematical equation in that it uses mathematical symbols, the symbols do not have their standard mathematical meaning. They represent (or are mapped onto) causal (and not solely mathematical) relations. If the model represented purely mathematical relations between the relevant variables, then it could also be satisfied by an excitatory circuit and, specifically, by a circuit that does not perform a divisive inhibition but that does multiplicatively excite a neural response. The purely mathematical relations between the variables of this excitatory circuit are equivalent to those described by a purely mathematical interpretation of the normalization model.

Purely mathematical descriptions have all the implications that can be arithmetically deduced. If the relations described by the mathematical interpretation of the normalization model can be deduced from the mathematical organization of an excitatory circuit, then the model does not describe some features that are relevant difference makers for COS. In some counterfactual situation in which the response of a simple cell in V1 is not decreased but rather is increased by its non-preferred stimulus (that is, some situation in which COS does not occur), the mathematical relations between the responses could remain the same as in the actual situation. We could describe the circuit in this counterfactual situation using the term "E_i" to refer to the simple cell response, the term "$\bar{E}_i$" to refer to its driving input, and the term "ΣE" to refer to a modulatory input that multiplicatively excites the simple cell. If this is the case, these three variables will still satisfy the relations described by the (mathematical interpretation of the) normalization equation (that is, $\bar{E}_i$ will still be equal to E_i divided by ΣE). This means that in a purely mathematical interpretation, the model could not be employed to address the w-question of what would happen if COS did not occur. As we saw in section 2, this is one of the kinds of w-questions that are relevant for CNCs (how would the computations change if the task were different). Therefore, if normalization is explanatory according to LB and Woodward's requirement, then it cannot describe merely mathematical relations, but rather causal relations and their quantitative properties.

The purely mathematical description can be contrasted with a description in which a mathematical "equation" represents a causal process that implements a mathematical function. In this description, the terms on one side of the equation refer to (are mapped onto) the *inputs* of the process, and the result on the other side refers to the *output*. If we interpret the normalization model this way, then this "equation" describes causal relations mathematically characterized. The model represents a causal relation (a divisive inhibition) between an output (the normalized response $\bar{E}_i$) and its inputs (the non-normalized response E_i and the sum of responses ΣE of all simple cells in the local population). To be satisfied by the excitatory circuit, the term $\bar{E}_i$ should represent one of its inputs and not the output, as it does in the normalization model under this causal interpretation. Also, ΣE should not appear as a denominator, which indicates an inhibitory modulation under this interpretation. The description of the computation would change if the behavior were excitatory. Therefore, this interpretation can be said to capture the difference makers of the relevant phenomena. This is why both LB and Woodward's criteria motivate a mechanistic interpretation of the modeling of neural arithmetic operations that define different CNCs.

One may worry that this is not sufficient to affirm that these models are mechanistic or, in other terms, that my argument presupposes conditions for abstract mechanistic explanation that are too weak, thus trivializing the notion of a mechanistic explanation. CNCs models describe the inputs and outputs of a neural system. But mere input–output descriptions are, according to mechanism, phenomenological models. Phenomenological models physically or mathematically represent the input–output behavior of a system without revealing anything about the underlying mechanisms, merely "saving" the explanandum phenomenon (Craver, 2006; Mauk, 2000). Mechanism implies that phenomenological models are not (fully) explanatory.

To provide a mechanistic explanation, a model must not only represent an input/output behavior but also satisfy the 3M requirement. A model that satisfies 3M to some degree is, by definition, not phenomenological. If a model describes some of the components and activities underlying an explanandum behavior, then it does not merely describe the input–output mapping that defines that behavior. CNC models satisfy 3M because the input–output mappings they describe are more complex than those required to describe their explananda. For instance, the normalization model includes more input variables than those required to represent COS. COS is a phenomenon that can be described by the variables that refer to the preferred stimulus, the non-preferred stimulus, and the suppressed neural response. The variable "Σ*E*" representing the modulatory input and the division symbol "–" that maps onto its suppressive activity are not part of the description of phenomenon. They refer to an inhibitory modulation that shapes the input–output relation defining the phenomenon.

Furthermore, the main functions that are attributed to neurons exhibiting COS, such as stimulus selectivity maximization, are not explained by normalization alone but rather by more complex computations. As we will see in the next section, normalization is only one computational step for the maximization of stimulus selectivity. A sum of inputs performed by a series of linear filters and an umbralization process are activities that, together with normalization, explain the relevant input–output behavior. More generally, explanation by neural arithmetic requires the specification of modulatory inputs (and their operations) that is not part of the input–output mapping used to define the phenomenon. This is why these models are not phenomenological.[5]

4. Mechanism and integration

The mechanistic interpretation of the normalization model not only accounts for its explanatory power but also for a further epistemic virtue. It has been argued that one main virtue of mechanism is that it can account for the increasing integration between different explanations produced by the development of cognitive neuroscience (Boone & Piccinini, 2016a). I claim that CNC models can be integrated with related models in different ways. They can be part of a hierarchy of mechanistic explanations of the kind described by Boone and Piccinini (2016a) and they can also take part in a horizontal integration of mechanistic models at the same degree of abstraction. This last kind of integration is what makes CNC models especially valuable for our understanding of neural processing, and this depends on the componential character of the mechanistic explanations in which (descriptions of) CNCs figure.

Boone and Piccinini (2016a) propose a mechanistic framework to understand the reconfiguration of cognitive science produced by the development of cognitive neuroscience. Traditional cognitive science was characterized by a division of labor between the study of a functional (or computational and algorithmic) level and the study of an implementation (neural or mechanistic) level. They consider that this framework is being replaced by cognitive neuroscience, where there is no longer a division of labor. According to the multilevel framework, mechanistic explanation of a phenomenon is given by a set of models that are not isolated but rather articulated in a hierarchy of mechanistic explanations. The description of activities (and organization) of components at a given level L_0 provides the explanation of activities of components at a subsequent superior level L_1. In turn, the description of activities (and organization) of components at L_1 provide the explanation for activities of components at a superior level L_2, and so on.

Models that describe canonical computations can be said to be integrated in this way with related models. The difference maker criterion employed in the previous section can motivate a hierarchy of mechanistic explanations in which CNC models can be embedded.[6] I will argue that the information excluded by the criterion as irrelevant for the explanandum of the normalization model refers or maps onto the difference makers for normalization itself. The description of neural circuits explains computations whose description explains informational tasks. For instance, there are specific features of shunting inhibition (one of the mechanisms that underlie normalization) that explain why the inhibition is specifically *divisive* (that is, they are difference makers for divisive inhibition).

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. Fast inhibitory transmission is typically mediated by $GABA_a$ (γ-aminobutyric acid type A) receptors, which conduct Cl– and HCO_3 ions and often have a reversal potential close to the resting potential. The increase in membrane conductance introduced by these synapses short-circuits the excitatory post-synaptic potentials (EPSPs) by locally reducing the input resistance. These shunting inhibitory conductances scale down EPSPs in a multiplicative manner in accordance with Ohm's law. Classical theoretical work (e.g., Blomfield, 1974; 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 (Figure 3). This means that there are features of shunting inhibition which, although they are not difference makers for stimulus selectivity (and therefore can be excluded from the normalization model), are indeed difference makers for normalization (and therefore must be included in the model that explains this behavior).

This implies that the difference maker criterion not only motivates the idea that CNC models provide mechanistic explanations but also that they take part in mechanistic *integration*. Information omitted from the computational explanation of informational tasks can be recycled by the model that explains the computations themselves. Mechanisms as shunting inhibition are part of a hierarchy in

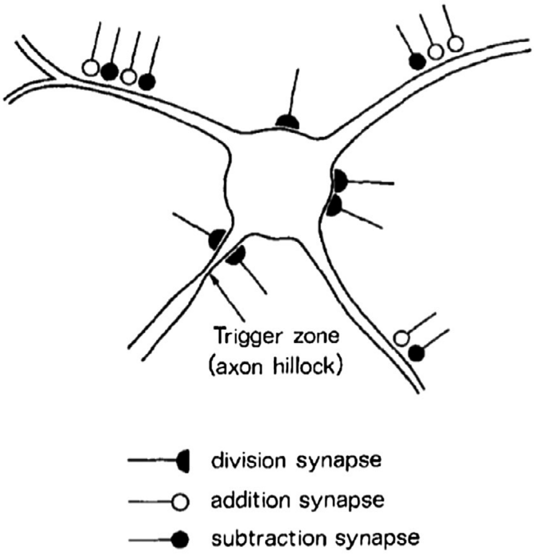

Figure 3. From Blomfield (1974). Divisive inhibition depends on the size and location of conductance.

which description of circuits explains computations whose description explains informational tasks.

The mechanistic approach to CNCs makes it possible for us to account for another kind of integration that I take to be more significant because, as mentioned earlier, it constitutes the special value that these models have for our understanding of neural processing. Levy and Bechtel (2013) emphasize the abstractness of Alon's models as a feature that is in contrast with a common way of understanding mechanistic models. Here, I want to emphasize another characteristic of these models, namely that they describe *motifs* or *patterns*, that is, abstract structures that are *recurrently* implemented in a system. Alon's motifs are connectivity patterns that occur in networks or biological circuits much more frequently than in random networks (Alon, 2007a, 2007b; Milo et al., 2002; Shen-Orr, Milo, Mangan, & Alon, 2002). One of Alon's contributions is precisely an algorithmic procedure to determine the presence of a motif in a biological network (Alon, 2007a, Chapter 3; Kashtan, Itzkovitz, Milo, Alon, 2004). The presence of recurrent components is a common feature in many artificial and biological systems. Electronic devices, for example, involve many instantiations of the same type of components at different levels (transistors, logic gates, memory registers, etc.). Alon maintains that biological systems, such as metabolic networks, also often employ the same component types. These components are modules in the sense of a set of nodes that interact and have a common function (Alon, 2003). Detection of these motifs can be employed to represent complex biological networks, such as the complete transcription network of E. coli, in a compact and modular way, where each component of the circuit can be identified with a different motif (Shen-Orr et al., 2002).

Alon (2003) points out that motif detection makes it possible not only to unveil the modular structure of a given circuit, but also to form a motif dictionary through which we can individuate the components of other circuits not yet studied. Understanding the modular structure of a new system facilitates the understanding of its global function. This means that motif detection makes it possible to unify different models at a methodological level since the same motif dictionary is used to model different mechanisms, and at an explanatory level, since the global functions of different systems can be explained by a common set of components or sub-functions.

The global functions of these systems can be completely different. This can be seen in the case of digital binary computation. The development of this kind of computation was born of the idea of employing a particular symbolic system to define a set of canonical abstract components for certain electrical circuits. In his *A Symbolic Analysis of Relay and Switching Circuits*, Shannon showed that Boolean or binary algebra can be used to analyze (and simplify the design of) the organization of relay circuits (the basic components of the electro-mechanic mechanisms employed by different communication devices at his time) (Shannon, 1938). Since then, the utilization of binary properties of electrical switches to perform logic functions is the basic concept underlying the design of most digital electronic computers.

In digital binary computation, we have two digits symbolically represented by "1" and "0," called *logic levels*. These levels can represent values of different physical variables. When digital binary computation is implemented by an electronic or electrical medium, the digits are value ranges of the voltage variable of a component of the system. The rules that regulate the digits of binary computation are a set of basic components defined by Boolean algebraic functions, such as conjunction, disjunction, and negation or inversion. These components, commonly referred to as logic gates, can be combined in various ways to form different mechanisms that implement very different global functions. For instance, sets of logic gates can be combined to build two combinatory (or non-sequential) circuits that perform different arithmetic operations: the addition and subtraction of digits. The addition of digits can be performed by a combination of logic gates known as a "full adder." This circuit consists of the combination of two circuits known as "half adders" that can be implemented by different combinations of logic gates. The combination of two half adders is necessary to have an additional input that works as a carry added to another column of higher power. In a similar manner, subtraction of digits is performed by a circuit known as a "full subtracter" composed of two "half subtracters," in order to have an additional input that works as a borrow from another column of higher power (Figure 4).[7]

Descriptions of different computational mechanisms explain different capacities by employing (the description of) a common set of basic operations. When we discover these basic operations through abstraction and employ them in different mechanistic models, I will say that a *horizontal* integration of these models is

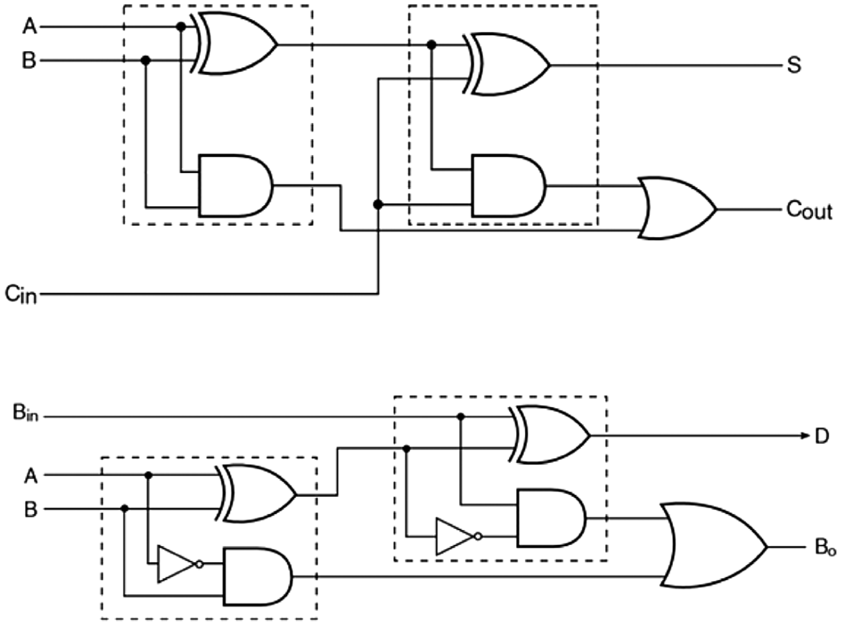

Figure 4. From Maini (2007). Arithmetic circuit diagrams for a full adder and a full substracter.

produced (as opposed to Boone and Piccinini's hierarchic or vertical integration). This integration is horizontal because it is the same degree of abstraction (the degree at in which the basic operations appear) that allows us to explain different capacities employing (the description of) a common set of basic operations. The integration is mechanistic because it depends on the compositional character of the relevant models. A set of models can be said to provide a unified understanding of different phenomena insofar as they describe different mechanisms constituted by (different arrangements of) a shared set of components (and their activities). In the case just considered, addition and subtraction of digits are different capacities that can be implemented by different mechanisms composed of logic gates.

I propose that the same kind of integration takes place in computational neuroscience as a result of the discovery of CNCs. As we saw in section 3, the neural computations considered by Carandini and Heeger are not merely abstract processes but also, like Alon's motifs, *canonical*. They are recurrently implemented by different systems of one organism and by different organisms. In fact, like Alon's motifs and logic gates, canonical computations can be combined in different ways to form more complex computations, which are responsible for different capacities.

In the previous section, we have seen that normalization is a computation relevant for the selectivity and invariance of V1 neurons. However, normalization does not produce this effect on its own, but rather is part of a more complex computational circuit. Divisive normalization operates on a sum of inputs that result from a series of linear filters and an umbralization process (e.g., Rust, Schwartz, Movshon, & Simoncelli, 2005). But normalization is also a component in a computational process related to attention. It is part of the computational process whose description explains the way in which a neural response is modified as a result of attention (Figure 5). In this model, normalization operates on the result of a multiplication performed by an attention field on the response of a neuron or population to its preferred stimulus (Reynolds & Heeger, 2009).

This example shows how descriptions of sets of canonical computations can be used to explain different cognitive processes or tasks in a unified way. In their defense of normalization as a canonical computation, Carandini and Heeger (2012) emphasize the value of this kind of unification. They argue that identifying and characterizing modular computations as normalization can provide a toolbox for obtaining a *principled* understanding of cerebral functions. They provide "a *unified* language to explain functional specialization of different brain areas" (p. 51).

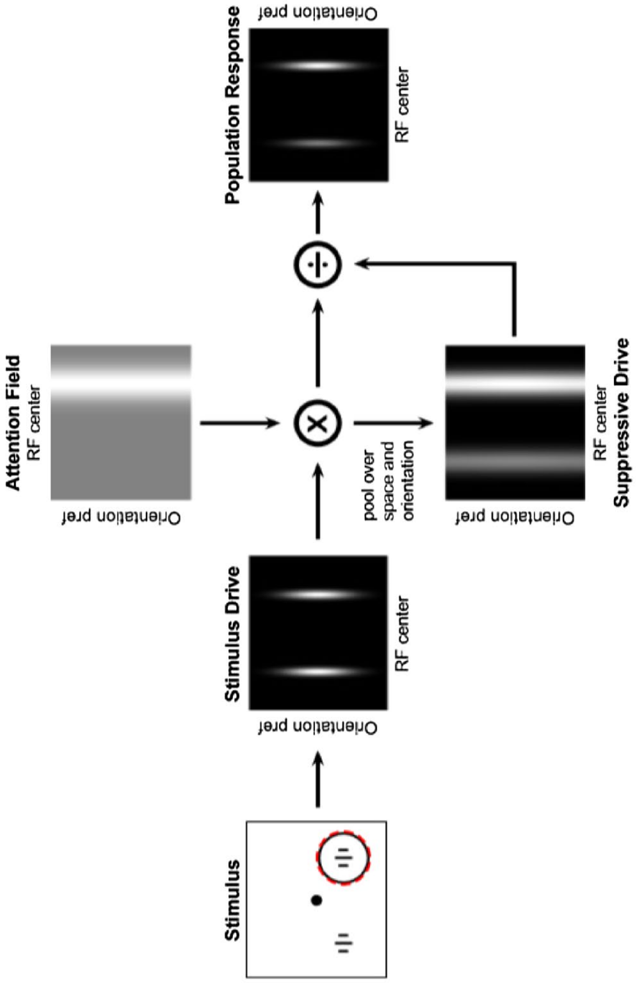

Figure 5. From Reynolds and Heeger (2009). Normalization operates on the result of a multiplication performed by an attention field on the response of a neuron to its preferred stimulus.

To conclude, it is useful to distinguish between the horizontal integration which I argue that abstract mechanistic models can provide and another epistemic virtue related to abstraction and emphasized by Levy and Bechtel (2013). The authors point out that abstraction in models such as Alon's (and also in CNCs' models) can maximize *generality.* Any mechanism that has the abstract connectivity described by Alon's model will exhibit the kind of behavior present in the bacteria arabinose system (namely, persistence detection). When this happens, abstraction implies a kind of unification in which apparently different phenomena or behaviors can be explained by the same model: a model that describes the same underlying abstract mechanism.[8] In contrast, the kind of unification that I am considering in this section occurs when different phenomena can be explained by *different* models that describe sets of *shared* or *common* basic components or operations. Abstract models, then, can provide an epistemic virtue in addition to generality maximization. It is a kind of theoretical integration which, like the hierarchical integration described by Boone and Piccinini, is non-reductive since it does not point to a privileged model at the expense of a variety of models, but integrates different models by providing a way to understand how they can be related to each other.

5. Conclusion

I have argued against the idea that canonical neural computation constitutes a counter-example for the thesis that mechanism can account for the explanatory power of abstract models. I claimed that the neurocognitive interpretation of these models is mechanistic in the sense that they describe the causal organization underlying a phenomenon and that only this interpretation (as opposed to a purely mathematical one) captures the relevant difference makers necessary for the model to be explanatory. I have also shown that mechanism is relevant to characterizing the kind of horizontal integration that makes CNC models especially valuable for our understanding of neural processing.

Notes

1. This approach to models is not universally accepted. Weisberg (2013) considers that we must distinguish between models and descriptions.
2. This notion is neutral regarding whether abstraction (or some specific form of abstraction) is compatible with explanation or whether abstraction diminishes the explanatory power of a model. An abstraction criterion for mechanistic explanation will be discussed shortly.
3. Alternatively, we can say that an abstract explanatory model can omit only the properties whose modification or replacement does not modify the *probability* of the behavior. This seems more accurate given that the relevant causes of a given behavior cannot be normally seen as sufficient to produce that behavior but only to increase its probability. I thank an anonymous referee for pointing out this alternative and more accurate version of the requirement. However, my main arguments do not depend

on this point. I will remain neutral regarding which version of the requirement we should endorse.

4. See, for example, Amos and James (2000, pp. 28–29).
5. I thank an anonymous referee for pointing out the relevance of clarifying that my argument does not imply a trivialization of the requirements for mechanistic explanation.
6. However, I do not claim that this hierarchy has the specific level structure that Boone and Piccinini propose.
7. See Mano (1979, pp. 116–119) and Maini (2007, Chapter 7).
8. As indicated at the end of the previous section, a mere input–output description does not count as a description of an abstract mechanism. A mechanistic model needs to satisfy 3M.

Acknowledgements

I am grateful to Professor William Bechtel for insightful comments that shaped the final version of this manuscript in significant ways, and also to the anonymous referees for many helpful suggestions. I am indebted to Liza Skidelsky, Sergio Barbieris, Sabrina Haimovici, Mariela Destéfano, Fernanda Velazquez, Nicolás Serrano, Magali La Rocca, Rodrigo Gonzáles Wilkens and Cristian Stábile for exhaustive discussions of this material.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science, 301*, 1866–1867.

Alon, U. (2007a). *An introduction to systems biology: Design principles of biological circuits*. Boca Raton, FL: Chapman & Hall.

Alon, U. (2007b). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics, 8*, 450–461.

Amos, S. W., & James, M. (2000). *Principles of transistor circuits: Introduction to the design of amplifiers, receivers, and digital circuits*. Oxford: Elsevier Science & Technology.

Barberis, S. (2013). Functional analyses, mechanistic explanations, and explanatory tradeoffs. *Journal of Cognitive Science, 14*, 229–251.

Bechtel, W., & 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 Research, 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., & Piccinini, G. (2016a). The cognitive neuroscience revolution. *Synthese, 193*, 1509–1534.

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

Carandini, M., & 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*, 127–153.

Chirimuuta, M. (in press). Explanation in computational neuroscience: Causal and non-causal. *British Journal for the Philosophy of Science*.

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*, 355–376.

Craver, C. F., & Kaplan, D. M. (2011). Towards a mechanistic philosophy of neuroscience: A mechanistic approach. In P. French & J. Saatsi (Eds.), *The continuum companion to the philosophy of science* (pp. 268–292). New York, NY: Continuum Publishing.

Haimovici, S. (2013). A problem for the mechanistic account of computation. *Journal of Cognitive Science, 14*, 151–181.

Heeger, D. J. (1992). Normalization of cell responses in the cat striate cortex. *Visual Neuroscience, 9*, 181–197.

Horowitz, P., & Winfield, H. (1989). *The art of electronics* (2nd ed.). Cambridge: Cambridge University Press.

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

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

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science, 78*, 601–627.

Kashtan, N., Itzkovitz, S., Milo, R., & Alon, U. (2004). Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs. *Bioinformatics, 20*, 1746–1758.

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science, 64*, 485–511.

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science, 80*, 241–261.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

Maini, A. K. (2007). *Digital electronics: Principles, devices and applications*. Hoboken, NJ: Wiley.

Mangan, S., Zaslaver, A., & Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology, 334*, 197–204.

Mano, M. M. (1979). *Digital logic and computer design*. Upper Saddle River, NJ: Prentice-Hall.

Mauk, M. D. (2000). The potential effectiveness of simulations versus phenomenological models. *Nature Neuroscience, 3*, 649–651.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science, 298*, 824–827.

Piccinini, G. (2015). *Physical computation, a mechanistic account*. Oxford: Oxford University Press.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron, 61*, 168–185.

Rice, C. (2015). Moving beyond causes: Optimality models and scientific explanation. *Noûs, 49*, 589–615.

Rust, N. C., Schwartz, O., Movshon, A. J., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron, 46*, 945–956.

Shannon, C. E. (1938). A symbolic analysis of relay and switching circuits. *Transaction of the American Institute of Electrical Engineers, 57*, 713–723.

Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics, 31*, 64–68.

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

Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.

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

Wainwright, M. J., Schwartz, O., & Simoncelli, E. (2002). Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons. In R. Rao, B. Olshausen, & M. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 203–222). Cambridge, MA: MIT Press.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford: Oxford University Press.

Woodward, J. F. (2003). *Making things happen*. New York, NY: Oxford University Press.

Woodward, J. F. (2013). Mechanistic explanation: Its scope and limits. *Proceedings of the Aristotelian Society, Supplementary, 87*, 39–65.

Woodward, J. F. (in press). Explanation in neurobiology: An interventionist perspective. In D. M. Kaplan (Ed.), *Integrating psychology and neuroscience: Prospects and problems*. Oxford: Oxford University Press.