Pre-referee eprint
Forthcoming in *Philosophy of Science*

Dynamical Models and Explanation in Neuroscience

Lauren N. Ross

Abstract

Kaplan and Craver claim that all explanations in neuroscience appeal to mechanisms. They extend this view to the use of mathematical models in neuroscience and propose a constraint such models must meet in order to be explanatory. I analyze a mathematical model used to provide explanations in dynamical systems neuroscience and indicate how this explanation cannot be accommodated by the mechanist framework. I argue that this explanation is well characterized by Batterman's account of minimal model explanations and that it demonstrates how relationships between explanatory models in neuroscience and the systems they represent is more complex than has been appreciated.

*To contact the author, please write to: Lauren N. Ross, Department of History and Philosophy of Science, University of Pittsburgh, 1017 Cathedral of Learning, 4200 Fifth Avenue, Pittsburgh, PA 15260; email: lnr18@pitt.edu.

†I would like to thank Robert Batterman, G. Bard Ermentrout, Mazviita Chirimuuta, Edouard Machery, Michael Miller, and James Woodward for helpful discussions and comments on earlier drafts of this paper.

Pre-referee eprint
Forthcoming in *Philosophy of Science*

1 Introduction

Recent philosophical discussion of explanation in the special sciences has focused on mechanist theories of explanation. These theories maintain that explanations appeal to the mechanisms that underlie the scientific phenomenon of interest. While there are different versions of these theories, a significant number of them define mechanisms as the entities, activities, and organizational features that produce a target phenomenon of interest.[1] This general mechanist position provides an explanatory structure that has been successfully identified in a number of special sciences, including neuroscience, molecular biology, and genetics. However, this success has led a number of philosophers to make the stronger claim that in certain sciences mechanistic explanation is the only form of explanation. One example of this stronger mechanist thesis is found in recent work by Kaplan and Craver who claim that all explanations in neuroscience are mechanistic (Kaplan and Craver 2011). They focus on dynamical systems neuroscience where there has been significant resistance to this strong mechanist position. This resistance has been motivated by claims that mathematical models in this field can provide explanations without referencing the mechanisms that underlie neural systems. In response to these claims Kaplan and Craver argue for two main points in regard to the explanatory status of mathematical models in dynamical systems neuroscience. They argue that these models (1) must meet a model-to-mechanism-mapping (3M) constraint to be explanatory and that (2) their explanatory status increases as they include more relevant mechanistic detail.

In this paper I argue against Kaplan and Craver's strong mechanist position and their claims regarding the explanatory status of mathematical models in dynamical systems neuroscience. I support this argument by analyzing a dynamical model that provides an explanation, despite failing to meet their mechanist requirements. I indicate how this explanation is well characterized by Batterman's account of minimal model explanations, which has been used to clarify the structure of explanations in the physical sciences, and more recently in biology (Batterman and Rice 2014). Understanding the explanation in this example involves attending to mathematical models and abstraction techniques that are common to dynamical systems neuroscience and used in understanding neural behavior. Such models can bear complex relationships to the neural systems they represent and although they do not meet the mechanist mapping requirements or represent the causal mechanical details of these systems, I show how this does not prevent them from providing explanations.

In the second section I describe Kaplan and Cravers mechanist position in

[1]Most mechanist theories that fall under this general account originate from (Machamer, Darden, and Craver 2000), while other versions can be found in (Bechtel and Richardson 2010; Glennan 1996; Woodward 2002).

more detail. The third section contains a brief background on dynamical systems neuroscience and an example of an explanation that these mathematical models provide. In the fourth section I argue that this type of explanation can not be accommodated by the mechanist approach and that it can be characterized by Battermans account of minimal model explanations. The sixth and final section contains a brief conclusion.

2 Kaplan and Craver's Mechanist Position

This section contains further description of Kaplan and Craver's claims regarding explanations in neuroscience, including their 3M constraint and the claim that the explanatory power of a model increases as it includes more mechanistic detail. They direct these claims at mathematical models in neuroscience and use them to distinguish between explanatory models and those that merely provide descriptions or predictions.

According to Kaplan and Craver, all explanations in neuroscience appeal to mechanisms as models in this field "carry explanatory force to the extent, and only to the extent, that they reveal (however dimly) aspects of the causal structure of a mechanism" (Kaplan and Craver 2011, 602). They define mechanisms as the underlying component parts of a system and the features, activities, and organization of these components that are relevant to the production of a particular phenomena of interest (Kaplan and Craver 2011, 605). Explaining this phenomenon requires citing all and only those actual components and activities that underlie and produce it. For example, an adequate explanation of neural firing (or the action potential) appeals to the relevant biological entities and activities that underlie and produce this firing. These biological entities include the relevant ion channels, ions, and the Na^+/K^+ pump, while the activities describe what these entities do, e.g. their attraction, blocking, diffusion, etc. (Craver 2008, 1025). As an account of causal explanation, the mechanist position depends on the rationale that explaining a phenomena of interest requires citing the causal factors that produce it. In other words, it requires that the explanans invoke factors that are causally relevant to the explanandum. If a model merely describes or predicts the explanandum, without citing the causal factors that produce it, the model is regarded as non-explanatory.

There are two central claims that Kaplan and Craver make regarding the explanatory status of mathematical models in neuroscience. The first is their model-to-mechanism-mapping (3M) constraint, which states:

> (3M) In successful explanatory models in cognitive and systems neuroscience (a) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of

> the target mechanism (Kaplan and Craver 2011, 611).

Although this statement of 3M explicitly addresses models in cognitive and systems neuroscience, Kaplan and Craver extend it to all models in neuroscience.[2] Their 3M constraint specifies two mapping-relations that must be met between the model and a target system in order for the model to be explanatory. The first maps the variables of a model to components within the system and the second maps dependencies among variables in the model to causal relations among components in the system. These criteria are intended to ensure that the model accurately represents the "internal aspects of the system" (Kaplan and Craver 2011, 616). However, the degree to which a model needs to fulfill 3M in order to be explanatory is not made entirely explicit in their work. They indicate that models need not completely map to the target system or refrain from idealizations and abstractions to be explanatory. Kaplan states that "3M requires only that *some* (at least one) of the variables in the model correspond to at least *some* (at least one) identifiable component parts and causal dependencies among components in the mechanism responsible for producing the target phenomenon" (Kaplan 2011, 347-8; emphasis original). In this manner, the 3M constraint is stated such that it requires only a minimal amount of mapping from the model to the target system.

The second main claim that Kaplan and Craver make is that among models meeting 3M, the explanatory power of a model increases as it includes more relevant mechanistic detail (Kaplan 2011, 347). According to Kaplan:

> As one incorporates more mechanistically relevant details into the model, for example, by including additional variables to represent additional mechanism components, by changing the relationships between variables to better reflect the causal dependencies among components, or by further adjusting the model parameters to fit more closely what is going on in the target mechanism, one correspondingly improves the quality of the explanation (Kaplan 2011, 347).

As including increasing amounts of detail into a model further reveals the causal structure of the mechanism, it increases the explanatory status of the model. Kaplan and Craver sometimes refer to this claim as a "fact" and at other times a "highly plausible assumption" (Kaplan and Craver 2011, 613; Kaplan 2011, 347). In either case, it is presented as a complement to their 3M constraint. This more-details claim provides a natural way of assessing the degree to which a model meets 3M or maps onto a causal mechanism. A more detailed mechanistic model, with a higher degree of mapping, will

[2]They focus on cognitive and systems neuroscience to argue that mechanistic explanation is the unique form of explanation in higher-level neuroscience, which they take to have already been established for lower-level neuroscience (Kaplan and Craver 2011, 602-3).

provide a better explanation because it will be able to answer a wider range of questions about the physical system of interest.

Kaplan and Craver provide a strong mechanist position in an ongoing debate about the explanatory status of dynamical models in neuroscience. They use their position to argue against the claim that dynamical models can be explanatory when they do not reveal the causal structure underlying system-level dynamics.[3] Dynamical models often contain variables representing macroscopic and behavioral features of neural systems and these variables do not always appear to map onto mechanisms in the 3M sense. In these cases, Kaplan and Craver claim that these variables "are not components in the sense of being the underlying parts of the mechanism" and merely provide mathematically compact characterizations of system-level behavior (Kaplan and Craver 2011, pp. 615-614). They state that these dynamical models provide at best *descriptions* or *predictions* of the behavior of complex mechanisms and that those who consider them explanatory "fundamentally misidentify the source of explanatory power in their models" (Kaplan and Craver 2011, 602). This criticism is directed towards those who have argued for distinctly dynamical, non-mechanistic explanations in neuroscience, which has been argued, most notably, by Stepp, Chemero, and Silberstein.[4] The most challenging objection these dynamicist accounts have faced is that non-mechanistic dynamical models are at best *descriptive* or predictive, but not explanatory. Unfortunately, these dynamicist arguments have remained susceptible to such an objection, because they have continued to reference the predictive success of these models without providing another clear sense in which they are explanatory.[5] The strong mechanist position has likely benefitted from the fact that these arguments for non-mechanistic dynamical explanation have not been viewed as entirely successful. Kaplan and Craver continue to uphold a strong mechanist position, whereby mechanistic considerations serve as the demarcation criterion between models that are explanatory and those that are non-explanatory.

I have described Kaplan and Cravers mechanist position and their claims regarding explanatory mathematical models in neuroscience. In the next section I describe a common type of mathematical model in neuroscience, a dynamical model, and how it is used to represent neural systems. I then

[3]In a separate paper, Kaplan argues for the 3M criteria in the context of computational neuroscience (Kaplan 2011). For an informative and helpful response to this paper and discussion of explanations in computational neuroscience see (Chirimuuta 2013).

[4]For these claims, see (Chemero and Silberstein 2008; Stepp, Chemero, and Turvey 2011, 12).

[5]In a very recent paper Chemero and Silberstein discuss specific challenges for the mechanistic strategies of localization and decomposition in neuroscience. They further distance themselves from their earlier predictivist claims and make a number of insightful preliminary claims regarding the structure of non-mechanistic dynamical explanations (Silberstein and Chemero 2013).

provide an example of how such a dynamical model is used to provide certain explanations in neuroscience.

3 Dynamical Systems Neuroscience

With this description of Kaplan and Craver's mechanist position, I move on to providing some background on dynamical systems neuroscience. In this section, I first discuss how neural excitably is understood and modeled with the dynamical systems approach. To do this I characterize neural excitably from a molecular perspective and contrast this with the dynamical systems perspective. After clarifying certain aims of dynamical modeling I provide an example of an explanatory dynamical model in neuroscience. I indicate why this dynamical model is explanatory and what led neuroscientists to seek the explanation it provides.

3.1 Dynamical models in neuroscience

In this section, I first discuss how neural excitably is understood and modeled with the dynamical systems approach. To do this I characterize neural excitably from a molecular perspective and contrast this with the dynamical systems perspective. After clarifying certain aims of dynamical modeling I provide an example of an explanatory dynamical model in neuroscience. I indicate why this dynamical model is explanatory and what led neuroscientists to seek the explanation it provides.

3.2 Dynamical Models in neuroscience

A major topic of study in neuroscience is the excitability of neurons as this is important for understanding how they transmit information. From the molecular perspective neural firing, or the action potential, is explained with a generic neuron model consisting of voltage-gated ion channels sensitive to Na^+ and K^+. When a neuron receives a strong enough signal a number of things happen in succession that cause it to fire. First, the sodium channels open quickly and Na^+ rushes into the cell causing the membrane potential to increase. This results in depolarization of the neuron and the upstroke of the action potential. Shortly after this depolarization, the potassium channels open and K^+ rushes out of the cell, while the sodium channels begin to close, decreasing the influx of Na^+. These events cause the membrane potential to decrease, which contributes to the repolarization of the neuron and downstroke of the action potential. The action potential travels down the length of the neuron and constitutes a single firing event.

In dynamical systems neuroscience, neural excitability is understood and modeled in a different way: the main aim is to study qualitative features of neural systems irrespective of their fine-grain molecular details. Qualitative features of neural systems are studied by analyzing the graphical and topological structures of dynamical models that represent these systems. A dynamical model is a mathematical model that describes how variables representing a particular system evolve with time. In neuroscience it is

common to model neural excitability in this way with coupled differential equations. For example, consider the following two-variable dynamical model:

$$\dot{V} = f(V, n) \tag{1}$$

$$\dot{n} = g(V, n) \tag{2}$$

This is a system of coupled differential equations that describe how V and n change over time. Here, V is the excitation variable, which represents neural factors responsible for depolarization, and n is the recovery variable, which represents neural factors responsible for repolarization. With this two-variable model the dynamical system can be represented graphically, as shown on the phase plane in Figure 1. In this figure V is plotted along the x-axis and n is plotted along the y-axis. To each point $(V, n) \in \mathbb{R}^2$ there is a corresponding vector whose x component is $\dot{V}$ and whose y component is $\dot{n}$. The vector field plotted shows $(\dot{V}, \dot{n})$ at each (V, n).

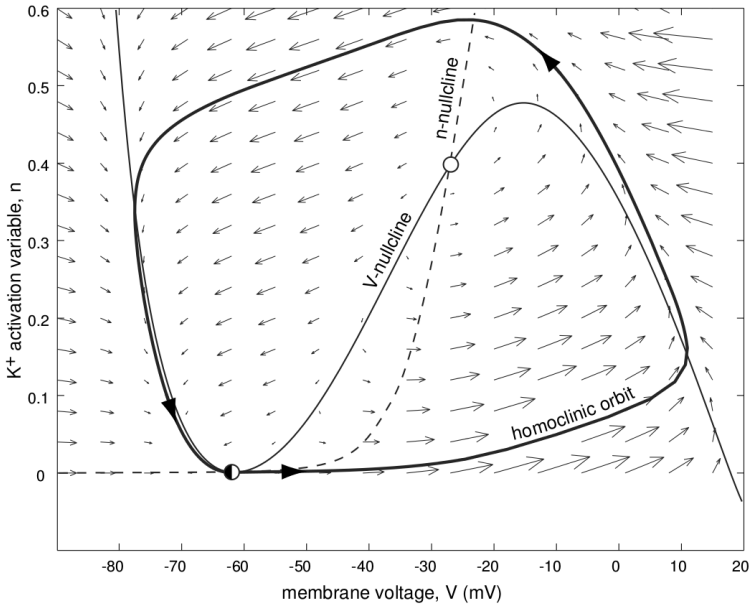

Figure 1: Phase plane with vector field (Izhikevich (2007), p. 113)

Graphical analysis of the vector field on the phase plane can provide information about the system that may not be obvious from the differential equations alone. For example, a solution to the system of equations can be obtained from an analysis of the figure, as it is the curve $(V(t), n(t))$ on the phase plane tangent to the vector field. The significance of a solution to the system of equations is that it gives a full picture of how V and n change over time. This solution and its portrayal as a trajectory corresponds to a characterization of neural firing. Counter-clockwise movement on the

trajectory tracks changes in V and n throughout the action potential and the completion of this trajectory represents a single firing of the neuron.[6]

In dynamical systems neuroscience the qualitative features of neural systems are often studied without reference to their fine-grained molecular detail. There are two main reasons for this. First, as graphical representations and qualitative features are exhibited by systems of differing molecular details, explaining these qualitative features does not depend on a shared physical structure. The fact that physically distinct neural systems can exhibit the same qualitative behavior motivates the view that this behavior is, in a sense, independent of any specific molecular microstructure. Hoppensteadt and Izhikevich express this sentiment when they state that "[b]ehavior can be quantitatively different, but qualitatively the same" (Hoppensteadt and Izhikevich 1997, 33). As the focus in dynamical systems neuroscience is on studying and explaining the qualitative behavior of neural systems, the physical differences among systems that exhibit these behaviors are rarely referenced (and sometimes the full extent of these differences are unknown). A second reason for this inattention to molecular detail is that preferred graphical analyses, which concisely represent the comprehensive behavior of neural systems, constrain the number of variables that can be implemented to characterize these systems. This requires the use of simple models that abstract from the molecular details of neural systems, while preserving their system-level behavior. The use of such techniques by Fitzhugh and Nagumo et al. in the early 1960s essentially marks the beginning of dynamical systems neuroscience (Fitzhugh 1960; Nagumo, Arimoto, and Yoshizawa 1962). Fitzhugh pioneered this work by reducing the number of variables in the Hodgkin-Huxley model of the action potential so that the system could be "easily visualized" in a phase-space, leading "to a better understanding of the complete system than can be obtained by considering all the variables at once" (Fitzhugh 1960, 873). He reduced the number of variables in these neural models by exploiting their different time scales and functional effects.[7] This early work explicitly distinguished the qualitative features of neurons and the topological properties of their phase space, from an analysis of their physical constitution. As I discuss in the following subsections, these techniques and the general aims of dynamical systems neuroscience are central to understanding how some models in this field are used to provide explanations.

3.3 Explanatory Dynamical Model: the Canonical Model

In this subsection, I give an example of a dynamical model in neuroscience and present an account of its role in a particular explanation. In this

[6]For more on graphical representations of neural excitability see (Ermentrout and Terman 2010; Izhikevich 2007).

[7]For Fitzhugh's use of these reduction techniques see (Fitzhugh 1960; Fitzhugh 1961) and for further discussion of them see (Abbott 1994; Doi and Kumagai 2001, 69).

example the dynamical model, referred to as a canonical model, represents the shared qualitative features of a number of physically distinct neural systems. I indicate how this dynamical model is used to provide explanations after discussing the research findings that led neuroscientists to seek these explanations.

In 1948 Hodgkin published important results from his voltage clamp studies of single crab neurons (Hodgkin 1948). In these experiments he measured the electrical responses of neurons after injecting them with various levels of current. He identified three different types of neural excitability, which he referred to as class I, class II, and class III excitability, a categorization still used today.[8] Class I neurons exhibit a low frequency of firing to low levels of current and smoothly increase their firing with increases in current. Class II neurons begin firing when the current stimuli reaches a higher level and their firing frequency increasing minimally with increases in current, as represented by the step function in Figure 2. The relationship between current introduced into class I and class II neurons and the frequency of their firing response is represented in the frequency-current (F-I) graph in Figure 2. Class III neurons fail to maintain firing in response to current stimuli (and are not depicted in the figure). The qualitative distinctions between these classes is that for class I neurons the frequency-current relationship starts at zero and increases continuously, for class II neurons it is discontinuous, and for class III neurons it is not defined.

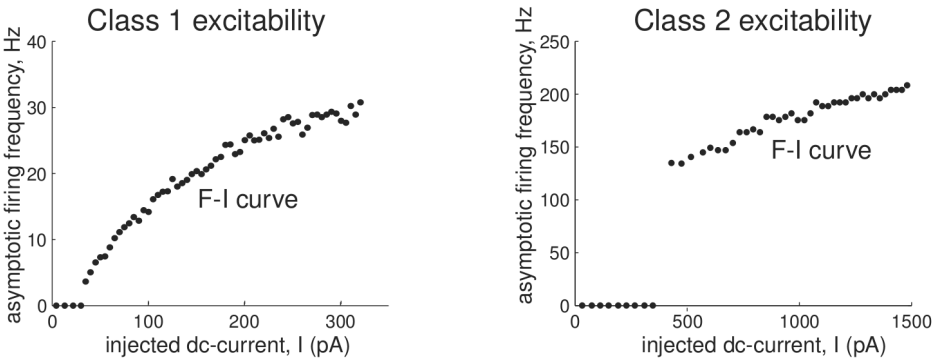

Figure 2: Graph of the frequency-current (F-I) relationship of class I and class II neurons (Izhikevich, p. 14 (2007))

These excitability classes identify qualitative features that are shared among large groups of physically distinct neurons. Hodgkin was particularly interested in class I excitability because it had been identified in neurons from many different animals (Hodgkin 1948, 167). Since his work, neuroscientists have identified class I excitability in many other neural

[8]These categories are sometimes referred to as type I, type II, and type III neuronal excitability (Hoppensteadt and Izhikevich 1997, 84).

systems, including rat hippocampal neurons, rat cortical neurons, crustacean motor neurons, and the majority of neurons in the mammalian cortex (Tateno 2004; Jia, Gu, and Li 2011; Connor 1975; Cauli, Audinat, Lambolez, Angulo, Ropert, Tsuzuki, Hestrin, and Rossier 1997). As neurons with class I excitability are found in animals of different biological phyla and even throughout the nervous systems of single species, it is unsurprising that this class encompasses neurons that differ in their microstructural details. What has been surprising, however, is the astonishing degree of this variation and the complexity of neural structures that has been revealed by recent advances in patch-clamp recording, heterologous expression of cloned channels, and genomic analysis (Bean 2007). For example, consider mammalian pyramidal neurons, the majority of which exhibit class I excitability. These neurons have three main types of voltage-gated ion channels responsible for excitability, including those selective for Na^+, K^+, and Ca^{2+}. Of those channels that transmit distinct ions each have an enormous variety of subtypes, for example, there are over 100 molecularly distinct K^+ channels (Vacher, Mohapatra, and Trimmer 2008). From this large selection of channels a single neuron typically expresses over a dozen different types, which vary in density along the neural membrane and result in many distinct voltage-dependent conductances. These voltage-dependent conductances contribute to the excitability of these neurons and can be comprised of 2-5 different currents each of ion (Na^+, K^+, and Ca^{2+}) (Bean 2007). This indicates a large degree of molecular difference among mammalian pyramidal neurons with class I excitability. The differences between all neurons that share this behavior is, of course, much greater.

Neuroscientists have sought an explanation for why neurons that differ so drastically in their microstructural details all exhibit the same type of excitability. In this case the explanadum is a behavior displayed by a group of physically distinct systems as opposed to a behavior produced by a single physically unique system. In 1986 Ermentrout and Kopell provided the crucial component of this explanation with their derivation of a canonical model for class I excitability.[9] Their work involves using mathematical abstraction techniques to reduce models of molecularly diverse neural systems to a single model, referred to as a canonical model. The canonical model and abstraction techniques used in this approach explain why molecularly diverse neural systems all exhibit the same qualitative behavior and why this behavior is captured in the canonical model. The explanation for this shared behavior is that when mathematical abstraction techniques are used to abstract away from details of mathematical representations of neural systems, all representations converge onto the same canonical model. In the next subsection, I further describe the abstraction steps, canonical

[9]This model is also called the "Ermentrout-Kopell model" and sometimes the "theta model" (Izhikevich 2004; Ermentrout, Rubin, and Osan 2002; Börgers, Epstein, and Kopell 2008).

model, and the explanations they provide.

3.3.1 Reducing models of neural excitability

The first step in this canonical model approach involves reducing the number of variables in models of neural excitability. Generally, variables characterizing the dynamics of neural systems are classified into four groups depending on their time scale and effect on membrane potential. These variables include: (1) the membrane potential variable, (2) excitation variables, (3) recovery variables, and (4) adaptation variables (Izhikevich 2007, 8). Excitation variables include neural factors that contribute to the upstroke of the action potential and firing of the neuron, while recovery variables represent neural factors that contribute to the downstroke of the action potential and recovery of the neuron. Adaptation variables stand for neural features that increase during continued spiking and can alter long-term neural excitably. This classification allows the factors characterizing neural excitability to be collapsed into one of the four variables that together characterize the dominant behaviors of the system.

Models for class I excitability do not contain variables of the fourth type, so our analysis begins with dynamical models characterized by three variables: the membrane potential variable, excitation variable, and recovery variable. A model with these three variables can be reduced to a two-variable model by exploiting differences in the rate of the kinetics between the excitation and recovery variables.[10] As the kinetics of the excitation variable are often much faster than the kinetics of the recovery variable, an idealization is introduced into the model by replacing the excitation variable with the value it quickly approaches (Rinzel and Ermentrout 1989). This reduces the model to two variables that characterize the macro-level behavior and dynamics of the system: the "new" excitation variable V, which was formerly the membrane potential variable,[11] and the recovery variable n. This two-variable dynamical model takes the same form as the coupled differential equations (1) and (2).

When models of neural excitability are reduced to two variables and represented graphically, those systems with class I excitability all exhibit the same change in topological structure as they transition from resting to sustained firing. This qualitative feature is captured in dynamical systems theory by the presence of a particular kind of bifurcation. In the case of neurons with class I excitability, all exhibit the saddle-node on invariant

[10]The use of scale differences to reduce variables in mathematical models is a well-known approach. For more on this approach, see: (Fowler 2007; Batterman 2000).

[11]Once this reduction is performed it is common to refer to the variable for the membrane potential as the "excitation variable." This is because the membrane potential variable tracks changes in the neural membrane due to current stimuli, which can result in excitation of the neural system.

circle bifurcation (Izhikevich 2007, 164). This reduction of mathematical models of neural excitability to two-variable models is the first step in the canonical model approach and begins to reveal the shared qualitative features in their topology.[12]

3.3.2 Ermentrout-Kopell Theorem

Identifying this particular bifurcation in all models of class I systems is significant because Ermentrout and Kopell's theorem for class I excitability proves that all models which exhibit this bifurcation transform into the same model when they are reduced. They prove this by providing a continuous piecewise transformation, represented by h in Figure 3, that transforms any one-variable model, among a family of models with this bifurcation, into a single canonical model.

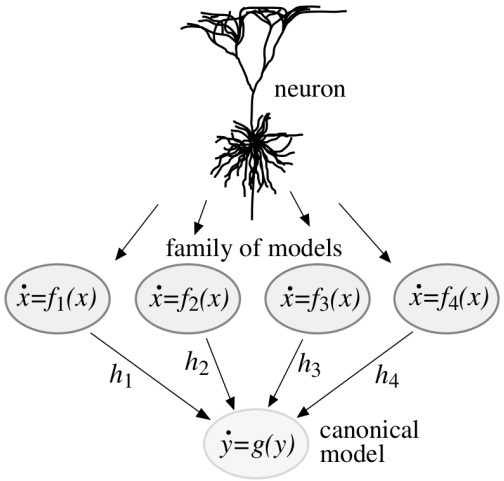

Figure 3: Modeling techniques in neuroscience (Izhikevich (2006))

In other words, Ermentrout and Kopell prove that all dynamical systems with the saddle-node on invariant circle bifurcation of the form:

[12]This first step allows for the representation of system-level behavior in a two-dimensional phase space and serves many important roles in understanding this behavior, e.g. in identifying the particular bifurcation that characterizes the system. However, for the purpose of reducing any model of neural excitability to the canonical model, so long as the system exhibits the saddle-node on invariant circle bifurcation, technically the Ermentrout-Kopell theorem is all that is required (Hoppensteadt and Izhikevich 1997).

$$\dot{x} = f(x), \quad x \in \mathbb{S}^1, \tag{3}$$

can be mathematically transformed into the following canonical model:[13]

$$\theta' = (1 - cos\theta) + (1 + cos\theta)r, \quad \theta \in \mathbb{S}^1, \tag{4}$$

where θ represents the activity of a neural system given a particular current input represented by r. Given a particular fixed value of the bifurcation parameter r, the model describes how the activity of the neural system, represented by θ, changes over time by specifying the location of θ on the unit circle $\mathbb{S}^1$. This is represented in Figure 4, where the location of θ on the unit circle indicates whether the neural system is in the rest, threshold potential, spike, or refractory phase. Every completion of the unit circle by θ represents a single firing event of the neural system. The model indicates that with small values of r the neural system remains at rest, represented by the variable θ at the rest potential position. Larger values of r result in continuous firing of the neural system, represented by the continuous movement of θ around the unit circle.

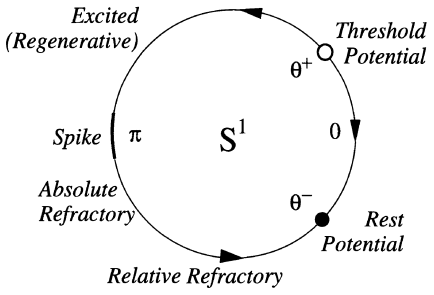

Figure 4: Physiological state diagram of a Class I neural system (Hoppensteadt & Izhikevich (1997), p. 228)

[13]Ermentrout and Kopell's theorem pertains not just to single neurons but also to neural networks. The equations that pertain to neural networks contain an extra term that accounts for the connectivity and interactions between neurons. For the these equations see (Hoppensteadt and Izhikevich 1997, 225). I have chosen the single neuron case for simplicity of presentation.

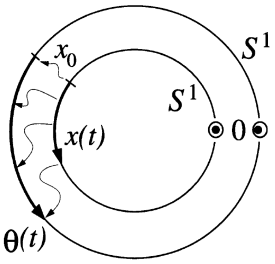

Figure 5: The solution $x(t)$ of (3) is mapped to the solution $\theta(t)$ of (4), the canonical model (Hoppensteadt & Izhikevich (1997), p. 119)

> Ermentrout and Kopell's theorem for class I excitability provides a continuous transformation $h : \mathbb{S}^1 \to \mathbb{S}^1$ that converts solutions of (3) to solutions of (4), represented in Figure 5. This figure shows how any point on the unit circle $\mathbb{S}^1$ of the family model is represented on the unit circle $\mathbb{S}^1$ of the canonical model. This transformation preserves the behavior of the original system and ensures that no artifacts or behavior not present in the original system are inherited by the canonical model. Neuroscientists describe this transformation as "extracting some particularly useful dynamical features" from these models, which are represented in the canonical model for class I excitability (Hoppensteadt and Izhikevich 1997, 115). This approach reveals how all models of systems with class I excitability are transformed into the same canonical model when they are reduced with principled mathematical techniques. This reveals how mathematical representations of class I systems are stable under certain perturbations by abstracting away from details of each model.
>
> One of the more impressive features of this canonical model is that it provides the frequency with which any class I neuron will oscillate given a particular fixed value of r (Hoppensteadt and Izhikevich 1997, 227-8). The canonical model approach is valued by mathematical neuroscientists because it provides a rigorous method for gaining information about classes of neural systems which share a particular behavior without obscuring this similarity behind the details of any one system (Izhikevich 2007, 278). As Izhikevich notes, the "advantage of this approach is that we can study universal neurocomputational properties that are shared by all members of the family because all such members can be put into the canonical form" (Izhikevich 2007, 278). Furthermore, as this canonical model approach pertains not just to single neurons, but also to neural networks, it indicates the relevance of this explanatory approach to both cellular and systems-level neuroscience.[14]
>
> It is worth emphasizing that this approach depends crucially on both the

[14]For more on Hoppensteadt and Izhikevich's discussion of the canonical modeling approach and its use in understanding weakly connected neural networks see (Hoppensteadt and Izhikevich 1997, 111).

canonical model and mathematical abstraction techniques that relate it to models of distinct neural systems. Referring to the canonical model alone could be viewed as merely describing or predicting the behavior of class I neurons, as opposed to explaining it. The canonical model approach, however, including the canonical model and abstraction techniques, does more than just describe or predict the excitability of class I neurons. It explains *why* physically distinct neural systems all share the same behavior by showing that principled mathematical abstraction techniques–which preserve qualitative behavior–can be used to reduce all models of these distinct systems to the same canonical model. These abstraction techniques involve exploiting time scale differences to introduce idealizations into models and transforming systems into simpler models that are topologically equivalent. This approach provides an explanation for this shared behavior – when principled mathematical techniques abstract from the details of different systems, they can all be simplified into the same canonical model that exhibits this behavior.

4 Analysis of the Canonical Model Approach

In this section, I examine whether Kaplan and Craver's 3M constraint and claims regarding detailed models–which they created to account for explanatory mathematical models in neuroscience–can accommodate the explanations provided by the canonical model approach. I argue that their mechanist criteria and framework cannot account for this type of explanation. I then describe Batterman's account of minimal model explanations and indicate how it accommodates the canonical model explanation introduced in the last section.

4.1 Kaplan and Craver's Mechanist Account

The canonical model approach contrasts with Kaplan and Craver's claims, because it is used to explain the shared behavior of neural systems without revealing their underlying causal mechanical structure. As the neural systems that share this behavior consist of differing causal mechanisms–different types of ion channels, with different distributions along the membrane, and permeabilities to specific ions, etc.–a mechanistic model that represented the causal structure of any single neural system would no longer represent the entire class of systems with this behavior. In other words, a mechanistic explanation can be provided to explain why any *single* system displays class I excitability, but this answers a different question than that answered by the canonical model, which takes the shared behavior of all systems in the class as the desired explanandum. As explaining this shared behavior is one of the goals of the canonical model, abstracting from these differences in mechanistic detail serves an explanatory purpose.
Furthermore, this explanatory purpose can contrast with the claim that the explanatory status of a model increases as more relevant mechanistic detail is included. This role of abstraction in dynamical systems neuroscience is

supported by a quote from Rinzel and Ermentrout, who state:

> [W]e emphasize the value of using idealized, but biophysically reasonable, models in order to capture the essence of system behavior. If models are more detailed than necessary, identification of critical elements is often obscured by too many possibilities. On the other hand, if justified by adequate biophysical data, more detailed models are valuable for quantitative comparison with experiments. The modeler should be mindful and appreciate of these two different approaches: which one is chosen depends on the types of questions being asked and how much is known about the underlying physiology (Rinzel and Ermentrout 1989).

To the extent that neuroscientists have sought to explain this universal behavior and have succeeded by purposefully abstracting from mechanistic detail, it should be regarded as a legitimate explanation. The canonical model approach indicates that there are common types of questions in dynamical systems neuroscience that are not answered by referencing causal mechanisms and often involve abstracting from many of these lower-level details.

These points can be made more clear by considering Kaplan and Craver's 3M constraint for explanatory models in neuroscience, which the canonical model does not meet. The first part of this constraint requires that the variables of a model map onto the mechanism of interest, i.e. the entities, activities, and organizational features of the target system producing the phenomena of interest. Recall that the canonical model contains a single variable θ and the bifurcation parameter r, representing the behavior of the neuron and a fixed input to the neuron, respectively. The bifurcation parameter does not represent a component (or internal aspect) of the neural system, but rather an input stimulation to the system. This leaves the variable θ as a candidate for the first part of the 3M constraint. This single variable (θ) cannot fulfill this constraint because it does not map onto any identifiable entity, activity, or organizational feature of the mechanisms that underlie these neural systems. Rather it represents the overall behavior of the neural system by indicating its location on the unit circle. The second 3M requirement–that variables in the model map onto causal relations in the target system–is also problematic. As the only candidates for a dependency relation in this model are θ and r, it may be claimed that they meet the second part of the 3M constraint: a dependency relation between a fixed input to the neuron and its behavior. However, the fact that these variables do not meet the first part of the 3M requirement makes this dependency relationship difficult to interpret with the mechanist framework. Furthermore, Craver considers this type of input/output relation to be a"phenomenal model" that "black boxes" the underlying causal mechanism. He claims that such phenomenal models are not explanatory because they

fail to represent the mechanism between the input and output relations. As he has stated, phenomenal models "are complete black boxes; they reveal nothing about the underlying mechanisms and so merely 'save the phenomenon' to be explained" (Craver 2006, 357). Thus the only possible dependency relation in the canonical model fails to meet 3M because it merely captures an input/output relation and fails to map onto an underlying causal structure.

Indicating that the canonical model does not meet the 3M constraint is not to say that the model does not represent or map onto neural systems in a manner relevant to its explanatory power. Surely models must bear some relationship to how things are in the real world in order to be explanatory. I am not arguing against there being an explanatorily relevant sense in which the canonical model maps onto physical systems. Instead I am arguing that the mechanists' 3M requirement does not accurately characterize this mapping relationship for all explanatory models in neuroscience. There does not seem to be any straightforward modification of 3M that would allow the mechanist to accommodate the complex relationship between the canonical model and systems with this type of shared behavior.

On this basis it is fair to conclude that the canonical model for class I excitably cannot be accommodated by Kaplan and Craver's mechanist account. This model fails to meet their 3M criteria, their claims regarding the inclusion of details in explanatory models, and their assertion that explanatory models reveal the structure of mechanisms. The specific example that I have provided indicates that even if the mechanist framework accounts for many explanations in neuroscience, it cannot not account for all of them.

4.2 Batterman's Minimal Model Explanations

An account of explanation that accommodates this canonical model example is Batterman's account of minimal model explanations. Explanations in science are often considered answers to why-questions and Batterman has distinguished between two different types of these questions: type (i) and type (ii) why-questions (Batterman 2001, 23). A type (i) why-question asks why a phenomenon manifests in a particular circumstance, while a type (ii) why-question asks why a phenomenon manifests generally or in a number of different circumstances. For example, a type (i) why-question might ask why a particular firing behavior is exhibited by a rat hippocampal neuron. An answer to this question is likely to provide an account of how components of the rat hippocampal neuron bring about the spiking behavior of interest. A type (ii) why-question, on the other hand, might ask why a particular firing pattern is found generally among a group of microstructurally distinct neurons, e.g. rat hippocampal neurons, crustacean motor neurons, and human cortical neurons. An answer to this question is unlikely to reference the lower-level components of the systems, because the components vary from system to system. An explanation for why all of these neurons exhibit the same firing behavior should explain why one can abstract away from the

details of each system to achieve the same higher-level behavior. Whenever the lower-level components of a single system are invoked, explanation of the shared behavior of all these systems is lost.

While mechanistic explanations provide answers to type (i) why-questions, Batterman's minimal model explanations aim to answer type (ii) why-questions. The first step in these explanations is the identification of a pattern or behavior that is shared among physically distinct systems. This shared behavior is often referred to as universal behavior and the group of systems that exhibit it as the universality class. The universality class can be delimited and made precise by using mathematical abstraction techniques to show how different physical systems display the same universal behavior. Batterman describes this strategy as involving an abstract space of possible systems, where each point in the space represents a particular physical system of interest.[15] The goal is to apply simplifying techniques to this space that allow for the elimination of details or degrees of freedom, while preserving the form of behavior of each system in the space. Repeated application of these techniques (which involve the renormalization group theory in Batterman's example) rescales the systems and changes their representation in a way that can be tracked as the movement of the system through this space. Studying the topological features of this abstract space reveal fixed points, or points in the space where many represented systems flow to and remain. Importantly, the systems in this space that flow to the same fixed point are in the same universality class and their shared behavior is determined by the fixed point that they all flow to. This procedure of creating, simplifying, and studying systems in this abstract space provides a precise way of delimiting the universality class (Batterman and Rice 2014). This strategy of delimiting a universality class explains why physically distinct systems all share the same behavior because it reveals that when details irrelevant to the behavior of each system are removed from the models that represent them, all systems share a common representation. As Batterman states:

> explanation of universal behavior involves the elucidation of principled reasons for bracketing (or setting aside as "explanatory noise") many of the microscopic details that genuinely distinguish one system from another. In other words, it is a method for extracting just those features of systems, viewed macroscopically, that are stable under perturbation of their microscopic details (Batterman 2001, 43).

Explaining this universal behavior answers a type (ii) why-question in explaining why physically distinct systems exhibit the same behavior.

[15]For more on Batterman's discussion of these points, see (Batterman 2001; Batterman 2010; Batterman and Rice 2014).

Delimiting the universality class can be used to identify what Batterman calls a "minimal model," which is known to be in the universality class and thus, shares features of all models in the class. A minimal model often provides a compact characterization of universal behavior and, as Nigel Goldenfeld states, is a model that "most economically caricatures the essential physics" (Goldenfeld, Martin, and Oono 1989; Batterman 2002). Thus, minimal models characterize the behavior of a universality class without representing the lower-level physical details of systems in the class. Such simple models are often used to study and explain universal behaviors, which Batterman refers to as minimal model explanations. What justifies the use of a minimal model in studying and explaining universal features? This justification is provided by the mathematical techniques that delimit the universality class and the identification of the minimal model as a member of this class.

There are striking similarities between Batterman's account of minimal model explanations and the explanations provided by the canonical model approach. Like minimal model explanations, the canonical model approach is used to explain the universal behavior of class I neurons. It provides an answer to a type (ii) why-question by explaining why a particular neural behavior is found among physically distinct neural systems. Models of these systems are represented in the abstract space of phase diagrams where mathematical techniques are used to identify the stable features of these models. As Hoppensteadt and Izhikevich write:

> "instead of saying that the [canonical] model loses information about the original phenomena, we say that our model is insensitive to the dynamics within an equivalence class...and that it captures properties [of models in the family] that are transversal to the partitioning" (Hoppensteadt and Izhikevich 1997, 116).

The canonical model for class I excitability is a minimal model in the sense that it provides a compact characterization of the behavior of a universality class, which has been precisely demarcated and includes the canonical model as a member. As Hoppensteadt and Izhikevich state:

> Canonical [m]odels arise when one studies critical regimes, such as bifurcations in brain dynamics. It is often the case that general systems at a critical regime can be transformed by a suitable change of variables to a canonical model that is usually simpler, but that captures the essence of the regime (Hoppensteadt and Izhikevich 1997, 4).

Moreover:

> Using comprehensive models [which attempt to take into account all known neurophysiological facts and data] could become a trap, since the more neurophysiological facts are taken into

> consideration during the construction of the model, the more sophisticated and complex the model becomes. As a result, such a model can quickly come to a point beyond reasonable analysis even with the help of a computer. Moreover, the model is still far from being complete (Hoppensteadt and Izhikevich 1997, 3, 5).

Mathematical neuroscientists abstract away from the physical differences among systems that exhibit class I excitability, in order to explain this shared behavior. This procedure involves extracting such behavior with mathematical reduction techniques and representing it with dynamical models. The dynamical models that concisely capture these shared behaviors are often referred to as canonical models. Neuroscientists consider the canonical model for class I excitability a "one-dimensional caricature of a 'real' neuron" and they use it to study and explain this universal neural behavior (Gutkin and Ermentrout 1998). An all too common objection to the explanatory status of dynamical models has been the claim that–in the absence of representing components of biological mechanisms–they are merely phenomenological models that are only capable of describing or predicting scientific phenomena. Kaplan and Craver insist that "there is no currently available and philosophically tenable sense of 'explanation' according to which such models explain," arguing that their mechanist theory alone best represents the standards of neuroscience. (Kaplan and Craver 2011, 602). This paper is intended to refutes such claims in light of Batterman's account of minimal model explanations and the similarity of this explanatory structure to explanations neuroscientists provide with the canonical model approach. This approach demonstrates how the techniques of dynamical systems neuroscience are used to explain *why* such universal behaviors are exhibited by physically distinct systems, as opposed to just providing descriptions or predictions of these behaviors or revealing their underlying causal mechanisms. Such explanations are provided by simplifying neural models of these systems in a way that reveals their shared qualitative features. That such features are represented by the canonical model is explained by using techniques to demarcate the universality class, of which the canonical model is a member.

I have indicated why Kaplan and Craver's mechanist position cannot account for the explanations provided by the canonical model approach and how they can be characterized by Batterman's account of minimal model explanations. This analysis indicates that there are explanations in neuroscience that do not meet Kaplan and Craver's mechanistic account of explanation and, thus, that it should not be considered the sole form of explanation in neuroscience.

5 Conclusion

I have argued that there are explanations in neuroscience that are not accommodated by Kaplan and Cravers mechanist theory of explanation. An

example of such an explanation is the canonical model approach, where a dynamical model is explanatory in virtue of abstracting from the physical details or mechanisms of distinct neural systems. I indicated how the explanatory structure of this approach can be characterized by Battermans account of minimal model explanations, which captures a role mathematical abstraction techniques can play in explaining universal behaviors. The canonical model approach shows how neuroscientists can understand and study neural behavior by deliberately removing details from models of these systems and that explanations in neuroscience can be attained even when simple mapping constraints are not met. It is typically presumed that for a model to be explanatory it must bear some relationship to how things are in the real world. The canonical model approach reveals that this relationship can be quite complex and that it is not captured by Kaplan and Cravers account of explanation in neuroscience.

References

Abbott, L. F. (1994). Single Neuron Dynamics: An Introduction. In F. Ventriglia (Ed.), *Neural Modeling and Neural Networks*, pp. 57–78. Pergamon Press.

Batterman, R. (2000). A 'Modern' (= Victorian?) Attitude Towards Scientific Understanding. *The Monist 83*, 228–257.

Batterman, R. and C. Rice (2014). Minimal Model Explanations. *Philosophy of Science 81*.

Batterman, R. W. (2001). *The Devil in the Details*. Asymptotic Reasoning in Explanation, Reduction, and Emergence. Oxford University Press, USA.

Batterman, R. W. (2002). Asymptotics and the Role of Minimal Models. *The British Journal for the Philosophy of Science 53*(1), 21–38.

Batterman, R. W. (2010). On the Explanatory Role of Mathematics in Empirical Science. *The British Journal for the Philosophy of Science 61*(1), 1–25.

Bean, B. P. (2007). The action potential in mammalian central neurons. *Nature Reviews Neuroscience 8*(6), 451–465.

Bechtel, W. and R. C. Richardson (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (2nd ed.). Princeton University Press.

Börgers, C., S. Epstein, and N. J. Kopell (2008). Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. *Proceedings of the National Academy of Sciences 105*(46), 18023–18028.

Cauli, B., E. Audinat, B. Lambolez, M. C. Angulo, N. Ropert, K. Tsuzuki, S. Hestrin, and J. Rossier (1997). Molecular and physiological diversity of cortical nonpyramidal cells. *The Journal of Neuroscience 17*(10), 3894–3906.

Chemero, A. and M. Silberstein (2008). Replacing Scholasticism with Science. *Philosophy of Science 75*(1), 1–27.

Chirimuuta, M. (2013). Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience. *Synthese*.

Connor, J. A. (1975). Neural repetitive firing: a comparative study of membrane properties of crustacean walking leg axons. *Journal of Neurophysiology 38*(4), 922–932.

Craver, C. F. (2006). When mechanistic models explain. *Synthese 153*(3), 355–376.

Craver, C. F. (2008). Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential. *Philosophy of Science 75*(5), 1022–1033.

Doi, S. and S. Kumagai (2001). Nonlinear Dynamics of Small-Scale Biophysical Neural Networks. In R. R. Poznanski (Ed.), *Biophysical Neural Networks: Foundations of Integrative Neuroscience*, pp. 261–302. Mary Ann Liebert, Inc. Publishers.

Ermentrout, B., J. Rubin, and R. Osan (2002). Regular traveling waves in a one-dimensional network of theta neurons. *SIAM Journal on Applied Mathematics 62*(4), 1197–1221.

Ermentrout, G. B. and D. H. Terman (2010). *Mathematical Foundations of Neuroscience*, Volume 35 of *Interdisciplinary Applied Mathematics*. Springer.

Fitzhugh, R. (1960). Thresholds and plateaus in the Hodgkin-Huxley nerve equations. *The Journal of General Physiology 43*(5), 867–896.

Fitzhugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal 1*(6), 445–466.

Fowler, A. C. (2007). *Mathematical Models in the Applied Sciences*. Cambridge University Press.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis 44*(1), 49–71.

Goldenfeld, N., O. Martin, and Y. Oono (1989). Intermediate Asymptotics and Renormalization Group Theory . *Scientific Computing 4*, 1–19.

Gutkin, B. S. and B. G. Ermentrout (1998). Dynamics of Membrane Excitability Determine Interspike Interval Variability: A Link Between Spike Generation Mechanisms and Cortical Spike Train Statistics. *Neural computation*, 1047–1065.

Hodgkin, A. L. (1948). The Local Elextric Changes Associated with Repetitive Action in Non-Medullated Axon. *Journal of Physiology* (107), 165–181.

Hoppensteadt, F. C. and E. M. Izhikevich (1997). *Weakly Connected Neural Networks*. Springer-Verlag New York Incorporated.

Izhikevich, E. M. (2004). Which Model to Use for Cortical Spiking Neurons? *IEEE Transactions on Neural Networks 15*(5), 1063–1070.

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience*. MIT Press (MA).

Jia, B., H.-G. Gu, and Y.-Y. Li (2011). Coherence-Resonance-Induced Neuronal Firing near a Saddle-Node and Homoclinic Bifurcation Corresponding to Type-I Excitability. *Chinese Physics Letters 28*(9), 090507.

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

Kaplan, D. M. and C. F. Craver (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science 78*(4), 601–627.

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

Nagumo, J., S. Arimoto, and S. Yoshizawa (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE 50*(10), 2061–2070.

Rinzel, J. and G. B. Ermentrout (1989). Analysis of neural excitability and oscillations. In *Methods in Neuronal Modelling: From synapses to Networks*, pp. 135–169. Cambridge, MA: MIT Press.

Silberstein, M. and A. Chemero (2013). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences. *Philosophy of Science 80*(5), 958–970.

Stepp, N., A. Chemero, and M. T. Turvey (2011). Philosophy for the Rest of Cognitive Science. *Topics in Cognitive Science 3*(2), 425–437.

Tateno, T. (2004). Threshold Firing Frequency-Current Relationships of Neurons in Rat Somatosensory Cortex: Type 1 and Type 2 Dynamics. *Journal of Neurophysiology 92*(4), 2283–2294.

Vacher, H., D. P. Mohapatra, and J. S. Trimmer (2008). Localization and Targeting of Voltage-Dependent Ion Channels in Mammalian Central Neurons. *Physiological Reviews 88*(4), 1407–1447.

Woodward, J. (2002). What Is a Mechanism? A Counterfactual Account. *Philosophy of Science 69*(S3), S366–S377.