To appear in a volume entitled *Integrating Psychology  and Neuroscience: Prospects and Problems*. (ed.  David Kaplan). Published version may be a bit different.

# Explanation in Neurobiology: An Interventionist Perspective[1]

## 1. Introduction

Issues about explanation in psychology and neurobiology have received a great deal of philosophical attention lately. To a significant degree this reflects the impact of discussions of mechanism and mechanistic explanation in recent philosophy of science. Several writers (hereafter mechanists), including perhaps most prominently, Carl Craver and David Kaplan (Craver 2000, 2006; Kaplan and Craver 2011, Kaplan 2011), have argued that at least in psychology and neuroscience, mechanistic theories or models are the predominant mode of explanation, with other sorts of theories or models often being merely "descriptive" or "phenomenological" rather than explanatory[2]. Other writers such as Chermero and Silberstein (2008) have disputed this, arguing that, e.g., dynamical systems models are not mechanistic but nonetheless explanatory. This literature raises a number of issues, which I propose to examine below. First, how should we understand the contrast between explanatory and descriptive or phenomenological models within the context of neuroscience? What qualifies a theory or model as "mechanistic" and are there reasons, connected to some (plausible) general account of explanation, for supposing that only mechanistic theories explain? Or do plausible general theories of explanation suggest that other theories besides mechanistic ones explain? In particular, what does a broadly interventionist account of causation and explanation suggest about this question? If there are plausible candidates for non-mechanistic forms of explanation in psychology or neurobiology, what might these look like? What should we think about the explanatory status of "higher level" psychological or neurobiological theories that abstract away from "lower level" physiological, neurobiological or molecular detail and are, at least in this respect, "non-mechanistic?"

In what follows I will argue for the following conclusions. First, I will suggest that an interventionist framework like that developed in Woodward (2003) can be used to distinguish theories and models that are explanatory from those that are merely descriptive. This framework can also be used to characterize a notion of a mechanistic explanation, according to which mechanistic explanations are those that meet interventionist criteria for successful explanation and certain additional constraints as well. However, from an interventionist perspective, although mechanistic theories have a number of virtues, it is a mistake to think that mechanistic models are the exclusive or

[1] Thanks to Mazviita Chirimuuta and David Kaplan for helpful comments on an earlier draft.

[2] David Kaplan has informed me that the intention in Kaplan and Craver, 2011 was not to exclude the possibility that there might be forms of non-mechanistic explanation that were different from the dynamical and other models the authors targeted as non-explanatory. At Kaplan's suggestion, I have adopted the formulation in this sentence (mechanism as "the predominant mode of explanation") to capture this point.

uniquely dominant mode of explanation in neuroscience and psychology. In particular, the idea that models that provide more mechanistically relevant low-level detail[3] are, even ceteris paribus, explanatorily superior to those which do not is misguided. Instead, my contrasting view, which I take to be supported by the interventionist account as well as modeling practice in neuroscience, is that many explanatory models in neurobiology will necessarily abstract away from such detail At the same time, however, I think that

[3] As Kaplan has observed in correspondence almost everyone agrees that the addition of true but irrelevant detail does not improve the quality of explanations; the real issue is what counts as "relevant detail" for improving the quality of an explanation. Kaplan (2011) thinks of relevant detail as a "*mechanistically* relevant detail" (my emphasis):

> 3M [Kaplan's and Craver's requirements on mechanistic explanation—see below] aligns with the highly plausible assumption that the more accurate and detailed the model is for a target system or phenomenon the better it explains that phenomenon, all other things being equal (for a contrasting view, see Batterman 2009). 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.

One possible understanding of "relevant detail" is detail about significant difference-makers for the explananda we are trying to explain—a detail is "relevant" if variations in that detail (within some suitable range) would "make a difference" for the explananda of interest (although possibly not for other explananda having to do with the behavior of the system at some other level of analysis). This is essentially the picture of explanation I advocate below. I take it, however, that this is probably not what Kaplan (and Craver) have in mind when the speak of mechanistically relevant detail, since they hold, for example, that the addition of information about the molecular details of the opening and closing of individual ion channels would improve the explanatory quality of the original Hodgkin-Huxley model even though (assuming my argument below is correct) this information does not describe difference-makers for the explanandum represented by the generation of the action potential. (This molecular information is difference-making information for *other* explananda.) Similarly, Kaplan differentiates his views from Batterman in the passage quoted above, presumably on the grounds that the information that Batterman thinks plays an explanatory role in, e.g., explanations of critical point behavior in terms of the renormalization group (see below), is not *mechanistically* relevant detail. So while it would be incorrect to describe Kaplan and Craver as holding that the addition of just any detail improves the quality of explanations, it seems to me that they do have a conception of the sort of detail that improves explanatory quality that contrasts with other possible positions, including my own (and Batterman's). I've tried to do justice to this difference by using the phrase "mechanistically relevant detail" to describe their position.

the mechanists are right, against some of their dynamicist critics, in holding that explanation is different from prediction (and from subsumption under a "covering law") and that some of the dynamical systems-based models touted in the recent literature are merely descriptive rather than explanatory. This is not, however, because all such dynamical systems models or all models that abstract away from implementation detail are unexplanatory, but rather because more specific features of some models of this sort render them explanatorily unsatisfactory.

The remainder of this chapter is organized as follows. Section 2 discusses some ideas from the neuroscience on the difference between explanatory and descriptive models. Sections 3 and 4 relate these ideas to the interventionist account of causation and explanation I defend elsewhere (Woodward, 2003). Section 5 discusses the idea that different causal or explanatory factors, often operating at different scales, will be appropriate for different models, depending on what we are trying to explain. Section 6 illustrates this with some neurobiological examples. Section 7 asks what makes an explanation distinctively "mechanistic" and argues that, in the light of previous sections, we should not expect all explanation in neuroscience to be mechanistic. Section 8 argues that, contrary to what some mechanists have claimed, abandoning the requirement that all explanation be mechanistic does not lead to instrumentalism or other similar sins. Section 9 illustrates the ideas in previous sections by reference to the Hodgkin-Huxley model of the generation of the action potential. Section 10 concludes the discussion.

## 2. Explanatory versus Descriptive Models in Neuroscience

Since the contrast between models or theories that explain and those that do not will be central to what follows, it is useful to begin with some remarks from some neuroscientists about how they understand this contrast. Here is a representative quotation from a recent textbook:

> The questions what, how, and why are addressed by descriptive, mechanistic, and interpretive models, each of which we discuss in the following chapters. Descriptive models summarize large amounts of experimental data compactly yet accurately, thereby characterizing what neurons and neural circuits do. These models may be based loosely on biophysical, anatomical, and physiological findings, but their primary purpose is to describe phenomena, not to explain them. Mechanistic models, on the other hand, address the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry. Such models often form a bridge between descriptive models couched at different levels. Interpretive models use computational and information-theoretic principles to explore the behavioral and cognitive significance of various aspects of nervous system function, addressing the question of why nervous systems operate as they do. (Dayan and Abbott, 2001)

In this passage, portions of which are also cited by Kaplan and Craver (2011), Dayan and Abbott draw a contrast between descriptive and mechanistic models, and suggest that the former are not (and by contrast, that the latter presumably are) explanatory. However, they also introduce, in portions of the above comments not quoted by Craver and Kaplan, a third category of model—interpretative models—which are also described as explaining (and as answering why questions, as opposed to the how questions answered by mechanistic models). The apparent implication is that although mechanistic models explain, other sorts of models that are not mechanistic do so as well, and both have a role to play in understanding the brain.

Dayan and Abbott go on to say, in remarks to which I will return to below, that:

> It is often difficult to identify the appropriate level of modeling for a particular problem. A frequent mistake is to assume that a more detailed model is necessarily superior. Because models act as bridges between levels of understanding, they must be detailed enough to make contact with the lower level yet simple enough to provide clear results at the higher level.

These remarks introduce a number of ideas that I discuss below: (1) Neuroscientists recognize a distinction between explanatory and merely descriptive theories and models[4]; (2) For purposes of explanation, more detail is not always better; (3) Different models may be appropriate at different "levels"[5] of understanding or analysis, with it often being

[4] One possible response to the use of words like "explanation", "understanding" and so on in these passages as well as those from Trappenberg immediately below, is that we should understand these words as mere honorifics, with the labeling of a theory as "explanatory" meaning nothing more than "I like it or regard it as impressive", rather than anything of any deeper methodological significance. It is not easy, however, to reconcile this suggestion with the care these authors take in contrasting explanatory models with those that are merely descriptive or phenomenological. Another more radical response would be to acknowledge that these authors to mean what they say but claim that they are simply mistaken about what constitutes an explanation in neuroscience with the correct view being the position advocated by mechanists. I assume, however, that few philosophers would favor such a dismissive response, especially since, as noted below, there are normative accounts of explanation (such as interventionism) which support the quoted ideas. Let me also add that although it is true that one motive for abstraction away from detail is to enhance computational tractability, the passages quoted and many of the examples discussed below make it clear that this is not the only motive: sometimes such abstraction leads to better explanations, where this is not just a matter of improved computational tractability.

[5] Talk of "levels" of explanation is ubiquitous in neuroscience, psychology, and philosophy, although many commentators (myself included—see Woodward, 2008) also complain about the unclarity of this notion. In order to avoid getting enmeshed in the philosophical literature on this subject, let me just say that the understanding of this notion I will adopt (which I think also fits with the apparent views of the neuroscientists discussed below) is a very deflationary one, according to which level talk is just a way of expressing claims about explanatory or causal relevance and irrelevance: To say that a multiple compartment model of the neuron (see section 6) is the right level for modeling dendritic currents (or an appropriate model at the level of such currents) is just to say that such a model captures the factors relevant to the explanation of dendritic currents. This gives us only a very local and contextual notion of level and also makes it entirely

far from obvious which level of modeling is most appropriate for a given set of phenomena; and (4) It is nonetheless important to be able to relate or connect models at different levels.

A second set of remarks come from a discussion of computational neuroscience modeling in Trappenberg (2002).

> As scientists, we want to find the roots of natural phenomena. The explanations we are seeking are usually deeper than merely parameterizing experimental data with specific functions. Most of the models in this book are intended to capture processes that are thought of as being the basis of the information-processing capabilities of the brain. This includes models of single neurons, networks of neurons, and specific architectures capturing brain organizations. ….
>
> The current state of neuroscience, often still exploratory in nature, frequently makes it difficult to find the right level of abstraction to properly investigate hypotheses. Some models in computational neuroscience have certainly been too abstract to justify claims derived from them. On the other hand, there is a great danger in keeping too many details that are not essential for the scientific argument. Models are intended to simplify experimental data, and thereby to identify which details of the biology are essential to explain particular aspects of a system.
>
> …. What we are looking for, at least in this book, is a better comprehension of brain mechanisms on explanatory levels. It is therefore important to learn about the art of *abstraction*, making suitable simplifications to a system without abolishing the important features we want to comprehend.

Here, as in the passage quoted from Dayan and Abbott, the notion of a finding an explanatory model is connected to finding the right "level" of "abstraction", with the suggestion that this has to do with discovering which features of a system are "essential" or necessary for the explanation of those phenomena. Elsewhere Trappenberg connects this to the notion of a "minimal" model— "minimal" in the sense that the model includes just those features or details which are necessary or required to account for whatever it is that we are trying to understand and nothing more[6]. Trappenberg writes that "we want the model to be as simple as possible while still capturing the main aspects of the data that the model should capture" and that " it can be advantageous to highlight the minimal features necessary to enable certain emergent properties in [neural] network [models]".

### 3. An Interventionist Account of Causation and Explanation

an empirical, aposteriori issue what level of theorizing is appropriate for understanding a given set of phenomena; it does not carry any suggestion that reality as a whole can be divided into "layers" of levels on the basis of size or compositional relations or that "upper level" causes (understood compositionally) cannot affect lower level causes.

[6] For recent discussions of the notion (or perhaps notions) of a minimal model see Chirimuuta, 2014 and Batterman and Rice, 2014.

How, if at all, might the ideas in these remarks be related to an interventionist account of causal explanation? I begin with a brief sketch of that account and then attempt to connect it to some issues about modeling and explanation in neuroscience suggested by the remarks quoted above. According to the interventional model, causal and causally explanatory claims are understood as claims about what would happen to the value of some variable under hypothetical manipulations (interventions[7]) on other variables. A causal claim of form *X* causes *Y* is true if (i) if some interventions that change the value of *X* are "possible" and (ii) under those interventions the value of *Y* would change. A more specific causal claim (e.g., that *X* and *Y* are causally related according to *Y=F(X)* where *F* is some specified function) will be true if, under interventions on *X*, *Y* responds in the way described by *F*. For our purposes, we may think of the following as a necessary condition for a structure *H* to count as a *causal explanation* of some explanandum *E*:

> *H* consists of true causal generalizations $\{G_i\}$ (true according to the criteria just specified) and additional true claims *C* (often but not always about the values taken by initial and boundary conditions) in the systems for which *H* holds such that $C \cup \{ G_i \}$ entails *E* and alternatives to *E* would hold according to $G_i$ if alternatives to *C* were to be realized (e.g. if those initial and boundary conditions were to take different values).

For example (cf. Woodward, 2003), an explanation of why the electromagnetic field due to presence of a uniform current along a long straight wire is given by the expression

$$(3.1) \quad E = 1/2\pi e_o L/r$$

(where *E* is the field intensity, *L* the charge density along the wire, and *r* the distance from the wire) might consist of a derivation of expression (3.1) from Coulomb's law, and facts about the geometry of the wire and the charge distribution along it, as well as information about how the expression describing the field would have been different if the geometry of the conductor or the charge distribution had been different, where (in this case) this will involve additional derivations also appealing to Coulomb's law. In this way the explanation answers a set of what Woodward, 2003 calls *what-if-things-had-been-different-questions*, identifying conditions under which alternatives to the explanandum would have occurred. This requirement that an explanation answer such questions is meant to capture the intuitive idea that a successful explanation should identify conditions that are explanatorily or causally *relevant* to the explanandum: the relevant factors are just those that "make a difference" to the explanandum in the sense that changes in these factors lead to changes in the explanandum. This requirement fits naturally with the notion of a minimal model on at least one construal of this notion: such a model will incorporate all and only those factors which are relevant to an explanandum in the sense described. The requirement also embodies the characteristic interventionist

[7] An intervention is an idealized, non-confounded experimental manipulation. See Woodward (2003).

idea that causally explanatory information is information that is in principle exploitable for manipulation and control. It is when this what-if things-had been different condition is satisfied that changing or manipulating the conditions cited in the explanans will change the explanandum. Finally, we may also think of this "what-if–things-had-been-different" condition as an attempt to capture the idea that successful explanations exhibit dependency relationships: exhibiting dependency relations is a matter of exhibiting how the explanandum would have been different under changes in the factors cited in the explanans.

Next a brief aside about non-casual forms of why explanations—another topic which I lack the space to discuss in the detail that it deserves. I agree that there are forms of why-explanation that are not naturally regarded as causal. One way of understanding these (and distinguishing them from causal explanations), defended in passing in Woodward, 2003, is to take causal explanations to involve dependency or difference-making relationships (that answer what-if-things-had-been- different questions) that have to do with what would happen under interventions. Non-causal forms of why-explanation also answer what-if- things-had-been-different questions but by citing dependency relations or information about difference-makers that does not have an interventionist interpretation. For example, the universal behavior of many systems near their critical point depends on certain features of their Hamiltonian but arguably this is not naturally regarded as a form of causal dependence—cf. footnote 10. The trajectory of an object moving along an inertial path depends on the affine structure of spacetime but again this is not plausibly viewed as a case of casual dependence. In what follows I will sometimes speak generically of dependency relations, where this is meant to cover both the possibility that these are causal and the possibility that they are non-causal.

Many different devices are employed in science to describe dependency relations between explanans and explanandum, including directed graphs of various sorts (with an arrow from *X* to *Y* meaning that *Y* depends in some way on *X*) Such graphs are widely used in the biological sciences). However, one of the most common (and precise) such devices involves the use of equations. These can provide interventionist information (or more generally information about dependency relations) by spelling out explicitly how changes in the values of one or more variables depend on changes (including changes due to interventions) in the values of others. In contrast to the tendency of some mechanists (e.g. Bogen, 2005) to downplay the significance of mathematical relationships in explanation, the interventionist framework instead sees mathematical relationships as playing a central role in many explanations, including many neuroscientifc explanations[8]. Often they are the best means we have of representing the dependency relations that are crucial to successful explanation.

In its emphasis on the role played by generalizations, including those taking a mathematical form, in explanation and causal analysis, the interventionist account has some affinities with the DN model. However, in other respects, it is fundamentally different. In particular, the interventionist account rejects the DN idea that subsumption under a "covering law" is sufficient for successful explanation; a derivation can provide

[8] This is certainly not true of all mechanists. Kaplan (2011) is a significant exception and Bechtel (e.g. Bechtel and Abrahamsen, 2013) has also emphasized the important role of mathematics in explanation in neuroscience and psychology.

such subsumption and yet fail to satisfy interventionist requirements on explanation, as a number of the examples discussed below illustrate. In addition, although the interventionist account requires information about dependency relations, generalizations and other sorts of descriptions that fall short of being laws can provide such information, so the interventionist account does not require laws for explanation. I stress this point because I want to separate the issue of whether the DN model is an adequate account of explanation (here I agree with mechanists in rejecting this model) from the issue of whether good explanations, including many in neuroscience, often take a mathematical or derivational form – a claim which I endorse. Interventionism provides a framework that allows for recognition of the role of mathematical structure in explanation without adopting the specific commitments of the DN model.

With these basic interventionist ideas in hand, now let me make explicit some additional features that will be relevant to the discussion below. First, in science we are usually interested in explaining regularities or recurrent patterns – what Bogen and Woodward (1988) call *phenomena* – rather than individual events. For example, we are usually interested in explaining why the field created by all long straight conductors with a uniform charge distribution is given by (3.1) rather than explaining why some particular conductor creates such a field. Or at least we interested in explaining the latter only insofar as the explanation we provide will also count as an explanation of the former. In other words, contrary to what some philosophical discussions of explanation suggest, it is wrong to think of explanation in science in terms of a "two stage" model in which one (i) first explains why some singular explanandum *E* (e.g. that a particular wire produces a certain field) by appealing to some low-level covering generalization *G* (e.g. 3.1) saying that *E* occurs regularly and then, in a second, independent step, (ii) explains why *G* itself holds via an appeal to some deeper generalization (e.g., Coulomb's law). Usually in scientific practice there is no separate step conforming to (i)[9]. Or, to put the point slightly differently, the low level generalization (*G*) is treated as something to be explained – a claim about a phenomenon – rather than as potential explainer of anything, despite the fact that many such *G*s (including (3.1)) qualify as "law-like", on at least some conceptions of scientific law.

Because claims about phenomena describe repeatable patterns they necessarily abstract away from some of the idiosyncrasies of particular events that fall under those patterns, providing instead more generic descriptions, often characterized as "stylized" or "prototypical". For example, the Hodgkin- Huxley model, described below, takes as its explanandum the shape of the action potential of an individual neuron, but this explanandum amounts to a generic representation of important features of the action potential rather than a description of any individual action potential in all of its idiosyncrasy. This in turn has implications for what an explanatory model of this explanandum should look like – what such a model aims to do is to describe the factors on which the generic features of this repeatable pattern depend, rather than to reproduce all of the feature of individual instances of the pattern. Put differently, since individual neurons will differ in many details, what we want is an account of how all neurons meeting certain general conditions are able to generate action potentials despite this variation.

[9] See Woodward, 1979 for additional argument in support of this claim.

This framework may also be used to capture *one* natural notion of a (merely) “phenomenological” model (but not the only one; see section 8 below): one may think of this as a model or representation that consists just of a generalization playing the role of *G* above – in other words, a model that merely describes some “phenomenon” understood as a recurrent pattern. Trappenberg (2002) provides an illustration[10]: the tuning curves of neurons in the LGN (lateral geniculate nucleus) may be described by means of class of functions called Gabor functions, which can be fitted to the experimental data with parameters estimated directly from that data. Trappenberg describes the resulting curves as a “phenomenological model” of the response fields in the LGN, adding that “ of course this phenomenological model does not tell us anything about the biophysical mechanisms underlying the formation of receptive fields and why the cells respond in this particular way” (p. 6). The tuning curves describe phenomena in the sense of Bogen and Woodward; they are generalizations which describe potential explananda but which are not themselves regarded as furnishing explanations. An “explanation” in this context would explain why these neurons have the response properties described by the tuning curves—that is, what these response properties depend on. Obviously, merely citing the fitted functions does not do this. As this example illustrates, this contrast between a merely phenomenological model and an explanatory one falls naturally out of the interventionist framework, as does the contrast between DN and interventionist conceptions of explanation. The fitted functions describe and predict neuronal responses (they show the neuronal responses to particular stimuli “were to be expected” and do so via subsumption under a “covering” generalization, which many philosophers are willing to regard as locally “lawlike” ), but they do not explain those responses on the interventionist account of explanation.

This idea that explanations are directed at explaining phenomena naturally suggests a second point. This is that what sorts of factors and generalizations it is appropriate to cite in an explanans (and in particular, the level of detail that is appropriate) depends on the explananda *E* we want to account for, where (remember) this will be characterization at a certain level of detail or abstractness. In providing an explanation we are looking for just those factors which make a difference to whatever explananda are our target, and thus it will be at least permissible (and perhaps desirable) not to include in our explanans those factors *S** which are such that variations or changes in those factors make no difference for whether *E* holds. (Of course, as illustrated below, an explanans that includes *S** may well furnish an explanation of some *other* explanandum *E** which is related to *E*—for example by describing the more detailed behavior of some particular set of instances of *E*.)[11]

[10] Kaplan (2011) also uses this illustration.

[11] There is a very large philosophical literature on abstraction, idealization, and the use of “fictions” in modeling which I will largely ignore for reasons of space. However, a few additional orienting remarks may be useful. First, a number of writers (e.g. Thomson-Jones, 2005) distinguish between *idealization*, understood as the introduction of false or fictional claims into a model, and *abstraction*, which involves omitting detail, but without introducing falsehoods or misrepresentation. I myself do not believe that thinking about the sorts of examples philosophers have in mind when they talk about “idealization” in terms of categories like “false” and “fictional” is very illuminating , but in any case it is

A physics example illustrates this point with particular vividness. Consider the "universal" behavior exhibited by a wide variety of different materials including fluids of different material composition and magnets near their critical points, with both being characterized by the same critical exponent *b*. In the case of fluids, for example, behavior near the critical point can be characterized in terms of an "order" parameter *S* given by the difference in densities between the liquid and vapor forms of the fluid $S = ó_{liq} - ó_{vap}$. As the temperature *T* of the system approaches the critical temperature $T_c$, *S* is found to depend upon a power of the "reduced" temperature $t= T-T_c/T$

$$S\sim |t|^b$$

Where *b* is the critical exponent referred to above. Remarkably, the same value of *b* characterizes not just different fluids but also the behavior of magnets in the transition from ferromagnetic to paramagnetic phases.

Suppose one is interested in explaining why some particular kind of fluid has the critical point that it does. Since different kinds of fluids have different critical points, the value of $T_c$ for any particular fluid will indeed depend on microphysical details about its material composition. However, if one is instead interested in explaining the universal behavior just described (the phenomenon or generic fact that $S \sim |t|^b$ with fixed *b* for many different materials), then (as particularly emphasized by Batterman in a series of papers—e.g. 2009) information about the differing microphysical details of different fluids is irrelevant: within the interventionist framework it is non-difference-making information. That is, this universal behavior does not depend on these microphysical details since, as we have just noted, variations in those details do not make a difference for whether this universal behavior occurs. In other words, the universality of this

worth emphasizing that the goal of including in one's model only those features that make a difference to some explanandum need not, in itself, involve the introduction of falsehood or misrepresentation; instead it involves the *omission* of non –difference-making detail. However, I will also add that I do not think that most of the cases of modeling of upper level systems discussed below are usefully viewed as involving *only* the omission of detail present in some lower level model—i.e. such upper level models do not just involve abstraction from a lower level model. Instead, such modeling typically introduces *new* detail/explanatory features not found in models of lower level systems—that is, it adds as well as removes. Of course if, like Strevens (2008), one begins with the idea that one has available a fundamental level theory *T* that somehow represents or contains "all" explanatorily relevant factors at all levels of analysis for a system (a neural "theory of everything") , then models of higher level behavior will involve only dropping various sorts of detail from *T*. But actual examples of lower level models in science are not like *T*—instead they include detail which is difference-making for some much more restricted set of explananda, with the consequence that when we wish to explain other higher level explananda, we must include additional difference-making factors. To take an example discussed in more detail below, one doesn't get the Hodgkin-Huxley model for the action potential just by omitting detail from a lower level multi-compartment model; instead the H-H model introduces a great deal of relevant information that is "new" with respect to any actual lower level model.

behavior shows us that its explanation must be found elsewhere than in details about the differences in material composition of different fluids. In fact, as Batterman argues, the explanation for universal behavior is provided by renormalization group techniques which in effect trace the behavior to very generic qualitative features (e.g., certain symmetries) that are shared by the Hamiltonians governing the interactions occurring in each of the systems, despite the fact these Hamiltonians differ in detail for each system[12].

This example provides a concrete illustration of the point made more abstractly by Abbot and Dayan and by Trappenberg: it is not always correct that adding additional accurate detail (for example, details about the different Hamiltonians governing the different systems above) improves the quality of one's explanation. Instead, this can detract from the goodness of the explanation if the target explanandum does not depend on the details in question. Or at the very least, it is not mandatory in constructing an explanation that one provide such detail. Arguably a similar point follows if the detail in question is "mechanistically relevant detail"—the explanatory import of the renormalization groups account of critical point behavior would not be improved by the provision of such detail.

### 4. "Levels" of explanation and independence

The general idea of an explanandum "not depending" on "lower level" or implementational/compositional/realizational detail deserves more development that I can give it here, but a few additional comments may be helpful in fleshing out the picture I have in mind. First, when we speak of non-dependence on such detail, what we have in mind is non-dependence within a certain range of variation of such detail, rather than complete independence from all facts about realization. For example, in the example discussed above, the value of the critical exponent *b* does not depend on variations in the composition of the fluid being investigated—whether it is water, liquid helium etc. This is *not* to say, however, that "lower-level facts" about such fluids play no role in determining the value of *b.* But the facts that are relevant are very generic features of the Hamiltonians characterizing these particular fluids – features that are common to a large range of fluids – rather than features that distinguish one fluid from another. To the extent there are materials that do not meet these generic conditions, the model will not apply to

[12] I gloss over a number of important issues here. But to avoid a possible misunderstanding let me say that the similarity between explanation of critical point behavior in terms of the renormalization group and the neurobiological explanations I consider is that in both cases certain behaviors are independent of variations in lower level details. However there is also an important difference: in the neurobiological cases, it often seems reasonable to regard the explanations as causal, in the case of the explanation of critical point behavior the explanation is (in my view and also in Batterman's) not causal. As suggested above, I would be inclined to trace this difference to the fact that in the neurobiological examples the explanatorily relevant factors are possible objects of intervention or manipulation. This is not the case for the renormalization group explanation. In this case, one can still talk of variations making or failing to make a difference, but "making a difference" should not be understood in causal or interventionist terms.

them. In a similar way, whether a relatively "high level" neural network model correctly describes, say, memory recall in some structure in the temporal lobe may be independent of various facts about the detailed workings of ion channels in the neurons involved in this structure—"independent" in the sense that the workings of these channels might have been different, within some range of variation (e.g., having to do with biologically normal possibilities), consistently with the network structure behaving in the same way with respect to phenomena having to do with memory recall. Again, this does not mean that the behavior of the structure will be independent of all lower level detail—for example, it certainly matters to the behavior of the network that the neurons are not made of copper wire or constituted in such a way that they disintegrate when connected. Just as with critical point behavior, the idea is that lower level facts about neuronal behavior will impose *constraints* on what is possible in terms of higher level behavior, but that these constraints often will be relatively generic in the sense that a number of different low level variants will satisfy them. In this respect what we have, is a picture involving, so to speak, partial or constrained autonomy of the behavior of upper level systems from lower level features of realization, but not complete autonomy or independence.

A second point worth making explicit is this: the picture just sketched requires that it be possible for a model or theory to explain *some* explananda having to do with *some* aspects of the behavior of a system without the model explaining explaining *all* such aspects. It is thus opposed to an alternative picture according to which to a theory that explains any explanandum satisfactorily must be a "theory of everything" that explains all aspects of the behavior of the system of interest, whatever the scale or level at which this is exhibited. In the neural case, for example, such a theory of everything would appeal to a single set of factors or principles that could be used to explain the detailed behavior of dendritic currents and ion channels in individual neurons, the overall behavior of large networks of neurons and everything in between. The alternative view which is implicit in the remarks from Dayan and Abbott and Trappenberg above is that in addition to being completely computationally intractable such a theory is not necessary to the extent that behavior at some levels does not depend on causal details at other levels. Instead, it is acceptable to operate with different models, each appropriate for explaining explananda at some level but not others. There will be constraint relationships among these models—they will not be completely independent of each other—but this is different from saying that our goal should be one big ur-model with maximal lower level detail encompassing everything[13].

[13] Two additional points: First, I do not mean to imply that "mechanists" like Kaplan and Craver are committed to such "a theory of everything" view. The point of my remarks above is just to make explicit some of the commitments of the picture I favor . Second, another way of putting matters is that on my view a model can, so to speak, designate a set of target explananda and say, in effect, that it is interested in explaining just these, rather than all behaviors at all scales exhibited by the system of interest. A model *M* that represents neurons as dimensionless points is, obviously, going to make radically false or no predictions concerning any phenomena *P* that depend on the fact that neurons are spatially extended, but it is legitimate for *M* to decline to take on the task of explaining *P*, if its target is some other set of explananda. In other words, *M* should be

## 5. The Separation of Levels/Scales

The ideas just described would be less interesting and consequential if it were not for another broadly empirical fact. In principle, it is certainly possible that a huge number of different factors might turn out, empirically, to make a difference (and perhaps roughly the "same" difference, if we were able to devise some appropriate measure for this) to some set of target explananda. It is thus of great interest (and prima-facie surprising, as well as extremely fortunate for modeling purposes) that this is often not the case. Instead, it often turns out that there is some relatively small number of factors that make a difference or at least a substantial or non-trivial difference to a target set of explananda. Or, to express the idea slightly differently, it often turns out that we can group or segregate sets of explananda in such a way that different sets can be accounted for by different small sets of difference-making factors. In physics, these sets (of explananda and their accompanying difference-makers) are sometimes described as "domains" or "regimes" or "protectorates" -- the idea being that certain explanatory factors and not others are "drivers" or represent the "dominant physics" for certain domains while other explanatory factors are the primary drivers for explananda in other domains. In physics, the possibility of separating domains and dominant explanatory factors in this way is often connected to differences in the "scale" (e.g., of length, time or energy) at which different factors are dominant or influential. That is, there often turn out to be factors that are very important to what happens physically at, say, very short length scales or at high energies but which we can entirely or largely ignore at longer length scales, where instead different factors (or at least factors characterized by different theories) become important. To take a very simple example, if we wish to understand what happens within an atomic nucleus, the strong and weak forces, which fall off very rapidly with distance are major determinants of many processes, and gravitational forces, which are very weak, are inconsequential. The opposite is true if one is interested in understanding the motion of galaxies, where gravity dominates. A similar point seems to hold for many biological phenomena, including phenomena involving the brain. Here too, considerations of scale – both temporal and length scale – seem to operate in such a way that certain factors are important to understanding phenomena at some scales and not others, while models appealing to other factors are relevant at other scales[14]. For example, the detailed behavior of ion channels in a neuron requires modeling at length and temporal scales that are several orders of magnitude less than is appropriate for models of the behavior of an entire neuron in generating an action potential. This suggests the possibility of models

assessed in terms of whether it succeeds in explaining the explananda in its target domain.

[14] One generic way in which this can happen is that factors that change very slowly with respect to the explananda of interest can be treated as effectively constant and hence (for some purposes) either ignored or modeled in a very simple way—by means of a single constant parameter. Another possibility is that some factor goes to equilibrium very quickly in comparison with the time scale of the explanandum of interest, in which case it may also be legitimate to treat it as constant.

that account for the latter without accounting for the former and vice-versa – a possibility described in more detail immediately below.

## 6. Levels of Modeling in Neurobiology

To illustrate the ideas in the preceding section in more detail, I turn to recent review paper entitled “Modeling Single-Neuron Dynamics and Computations: A Balance of Detail and Abstraction” (Herz et al. 2006). In this paper, the authors describe five different “levels” (there’s that word again) of single neuron modeling. At “level one” are “detailed compartment models” (in some cases consisting of more than 1000 compartments[15]) which are “morphologically realistic” and “ focus on how the spatial structure of a neuron contributes to its dynamics and function”. The authors add, however, that “[a]lthough detailed compartmental models can approximate the dynamics of single neurons quite well, they suffer from several drawbacks. Their high dimensionality and intricate structure rule out any mathematical understanding of their emergent properties.” By contrast, “reduced [compartment] models [level two] with only one or few dendritic compartments overcome these problems and are often sufficient to understand somatodendritic interactions that govern spiking or bursting”. They add that “a well-matched task for such [reduced compartment] models is to relate behaviorally relevant computations on various time scales to salient features of neural structure and dynamics”, mentioning in this connection the modeling of binaural neurons in the auditory brainstem.

Level three comprises “single compartment models” with the Hodgkin-Huxley model being explicitly cited as an example. Herz et al. write:

> Single-compartment models such as the classic Hodgkin-Huxley model neglect the neuron’s spatial structure and focus entirely on how its various ionic currents contribute to subthreshold behavior and spike generation. These models have led to a quantitative understanding of many dynamical phenomena including phasic spiking, bursting, and spike-frequency adaptation (p. 82)

They add that models in this class “explain why, for example, some neurons resemble integrate-and-fire elements or why the membrane potential of others oscillates in response to current injections enabling a ‘‘resonate-and-fire’’ behavior”, as well as other explananda (p. 82).

Cascade models (level four) involving linear filters, non-linear transformations and explicit modeling of noise abstract even further from physiological details but “allow one to capture additional neural characteristics” such as those involved in adaptation to light intensity and contrast. Finally, “black box models” (level five) which may

[15] “Compartment” refers to the number of sections, represented by distinct sets of variables, into which the neuron is divided for modeling purposes—for example, the HH model is a “single compartment” model since the modeling is in terms of a single variable, voltage, which characterizes the behavior of the entire neural membrane. A multiple compartment model would have many different voltage variables for different parts of the membrane.

characterize the behavior of a neuron simply in terms of a probability distribution governing its an input/out relationships may be most appropriate if we “want to understand and quantify the signal-processing capabilities of a single neuron without considering its biophysical machinery. This approach may reveal general principles that explain, for example, where neurons place their operating points and how they alter their responses when the input statistics are modified.” (p. 83) Models at this level may be used to show, for example, how individual neurons shift their input-output curves in such a way as to achieve efficient coding.

Several features of this discussion are worth particular emphasis. First, and most obviously there is explicit countenancing of models at number of “levels”, where the notion of level is tied to differences in spatial and temporal scale (a representation of the neuron as spatially extended, with different potentials in different spatial regions is required for understanding dendritic currents, but this scale of spatial representation may be not required for other purposes). Models at each level are explicitly recognized as being capable of providing “explanations”, “understanding” and the like, rather than models at some levels being regarded as merely descriptive or phenomenological in a way that contrasts with the genuinely “explanatory” models at other (presumably “lower”) levels. Moreover, these models are seen as complementary rather than in competition with each other, at least in part because they are seen aiming at different sets of explananda. There is no suggestion that we have to choose between modeling at a very fine-grained, detailed level (e.g., level one) or a more coarse-grained level (e.g., level four or five). Second, it is also recognized that which modeling level is most appropriate depends on the phenomena one wants to explain and that is not true that models with more details (or even more mechanistically relevant details) are always better, regardless of what one is trying to explain, although for some purposes highly detailed models are just what is called for[16]. For example, if one’s goal is to understand how the details of the anatomy and spatial structure of an individual neuron influence its detailed dynamics, a model at level one may be most appropriate. If one wants a “quantitative understanding” of spike train behavior, a model at a higher level (e.g., level three) may be better. This would be better in the sense that the details invoked in a level one model may be such that they are irrelevant to (make no difference for) this phenomenon. Again, the goal is taken to be the inclusion of just enough detail to account for what it is one is trying to explain but not more:

> All these [modeling] tasks require a delicate balance between incorporating sufficient details to account for complex single-cell dynamics and reducing this complexity to the essential characteristics to make a model tractable. The appropriate level of description depends on the particular goal of the model. Indeed, finding the best abstraction level is often the key to success. (p. 80)

## 7. Mechanistic Explanation

[16] Once again, my goal in these remarks is the positive one of highlighting a feature of good explanatory practice in neuroscience. I do not mean to imply that mechanistic approaches are unable to incorporate this feature, but rather to emphasize that they should.

So far I have discussed "explanation" but have said nothing about distinctively "mechanistic" explanations and how these relate to the ideas just described. Although, for reasons that will emerge below, I don't think that "mechanistic explanation" is a notion with sharp boundaries, I fully agree that these are *one* important variety of explanation in many areas of biology and neuroscience. Roughly speaking, I see these as explanations meeting certain specific conditions M (described immediately below) that lead us to think of them as "mechanistic", where satisfying M is one way of meeting the general interventionist conditions on explanation. However, I also think that it is possible for a theory or model to fail to satisfy conditions M and still qualify as explanatory in virtue of meeting these more general conditions.

At the level of methodology, if not underlying metaphysics, my general picture of mechanisms and mechanistic explanation is fairly close to that advanced by other writers, such as Machamer, Darden and Craver (2000) and Bechtel and Abrahamsen (2005). Consider a system *S* that exhibits behavior *B* – the phenomenon we want to explain. A mechanistic explanation involves decomposing *S* into components or parts ("entities" in the parlance of Machamer, Darden and Craver (2000)), which exhibit characteristic patterns of causal interaction with one another, describable by generalizations $G_i$ (describing "activities"). Explanation then proceeds by showing how *B* results from these interactions, in a way that satisfies the interventionist conditions on causal explanation. This in turn involves showing how variations or changes in the parts or in the generalizations governing them would result in alternatives to *B*, thereby allowing us to see how the behaviors of the parts and the way in which they interact make a difference for (or are relevant to) whether *B* holds. Part of the attraction of explanations that are mechanistic in this sense is that this information about the parts and their interactions can guide more fine-grained interventions that might affect behavior *B* – a point that is spelled out in detail in Woodward (2002) and Kaplan and Craver (2011).

Explanations having this general character often, and perhaps even typically, satisfy several other related conditions. One of these, which I have discussed elsewhere (Woodward 2003) is a *modularity* condition: modularity requires that the different causal generalizations $G_i$ describing the causal relations among the parts should at least to some degree be capable of changing independently of each other. Versions of modularity are often explicitly or implicitly assumed in the "box (or node) and arrow" representations that are adopted in many different disciplines for the representation of mechanisms, with modularity corresponding to the idea that arrows into one node can be disrupted without disrupting arrows into other nodes. Arguably, satisfaction of a modularity condition is also required if we are to make sense of the idea that mechanistic explanation involves decomposition of *S* into distinct "parts" with distinctive generalizations characterizing the behavior of parts and the interactions into which they enter. If the alleged parts can't be changed or modified (at least in principle) independently of each other or if no local changes can affect the pattern of interaction of some of the parts without holistically altering all of the parts and their interactions, then talk of decomposing the behavior of the system into interactions among its "parts" seems at best metaphorical. In practice, the most straightforward cases in which modularity conditions are satisfied seem to be those in which a mechanical explanation provides information about *spatio-temporally* separate parts and their spatio-temporal relations, since distinctness of spatio-temporal location is very closely tied to the possibility of independent modifiability. For example, the spatio-temporal separation of the different classes of ion channels (Na and K channels) in the Hodgkin-Huxley model discussed in section 9 is one reason why it is natural to think of that model as involving a representation of independently modifiable parts that interact to produce the action potential and thus to think of the HH model as in this respect a "mechanical" model[17].

A second feature possessed by explanations that we most readily regard as mechanistic (or at least a feature that, reasonably enough, philosophers favorable to mechanism often take to be characteristic of mechanistic explanations) is a kind of *sensitivity* of behavior to details (material and organizational) of implementation/ realization/composition. Consider some ordinary machine (e.g., a clock). For such a machine to function as it was designed to, these components must be connected up to one another in a relatively spatio-temporally precise way. Moreover, the details of the behavior of the parts also matter – we do not expect to be able to replace a gear in a clock with a gear of different size or different spacing to teeth and get the same result. Indeed, this is why we need to invoke such details to explain the behavior of these systems: the details make a difference for how such systems behave. It is systems of this sort for which "mechanistic" explanation (or at least the kind of mechanistic explanation that invokes considerable implementational detail) seems particularly appropriate.[18]

Putting these requirements together, we get the claim that mechanical explanations are those that satisfy the interventionist requirements in section 2, which involve decomposition into parts (where the notion of part is usually understood spatio-temporally), and which are appropriate to systems whose behavior is sensitive to details of material realization and organization. Since satisfaction of this last condition, in

[17] My claim here is that modularity and decomposition into independently changeable parts are conditions that are most readily satisfied when "part" is understood in spatio-temporal terms, but for purposes of this paper, I leave open the question of whether decomposition (and hence mechanistic explanation) might also be understood in a way that does not require spatio-temporal localizability of parts. (Bechtel and Richardson (1993) were among the first to talk about this kind of decomposition, which they called functional decomposition.) Cognitive psychology employs a number of different strategies that seek to decompose overall cognitive processes into distinct cognitive processes, components or modules (e.g., Sternberg 2001), but typically without providing information about the spatial location of those parts, although usually there is appeal to information about temporal relationships. Assessment of these strategies is beyond the scope of this paper, although I will say that the strategies require strong empirical background assumptions and that proposals about decompositions of cognitive processes into components often face severe under-determination problems in the absence of information about neural realization (which does provide relevant spatial information). (See also Piccinini and Craver 2011 for a discussion of closely related issues.]

[18] These features of sensitivity to details of organization and composition as characteristic of mechanical explanation are also emphasized in Levy (forthcoming) and in Levy and Bechtel (forthcoming). Woodward (2008) also distinguishes systems that are realization sensitive from those that are not, although not in the context of a discussion of mechanistic explanation.

particular, is a matter of degree, we should not expect sharp boundaries between mechanistic and non-mechanistic forms of explanation, although there will be clear enough cases. (The absence of such sharp boundaries is itself one reason for thinking that is misguided to suppose that only theories meeting mechanistic constraints explain—the notion of mechanisticl explanation is not sufficiently sharply bounded to play this sort of demarcational role.)

We have already noted many cases in which, in contrast to the mechanistic possibilities just described, we need to invoke only very limited information about the details of material realization or spatio-temporal organization to explain aspects of the behavior of a system. For example, the explanation of universal behavior near critical points in terms of the renormalization group does not appeal to the details of the composition of the particular materials involved, for the very good reason that such behavior does not depend on these details. In part for this reason, it seems unintuitive to describe the renormalization group explanation as a "mechanistic". Certainly it is not mechanistic in the sense of that notion employed by writers like Craver. Nonetheless the renormalization group analysis seems explanatory. Previous sections have also noted the existence of many "higher level" explanatory neurobiological models and theories that abstract away from many neural details. To the extent such models are relatively insensitive to material or organizational details of implementation or to the extent they do not involve decomposition of the system modeled into distinct parts with characteristic patterns of interaction, the models will seem also seem comparatively less mechanistic.

As an additional illustration, consider the very common use of models involving recurrent networks with auto-associative features to explain phenomena like retrieval of memories from partial cues. Such models represent neurons (or perhaps even populations of neurons) as individual nodes, the connections of which form directed cycles, with every node being connected to every other node in a fully recurrent network. In a separate training phase, the network produces, via a process of Hebbian learning, an output which resembles (imperfectly) some previously acquired trained pattern. This output is then fed back into the network, resulting in a pattern that is closer to the trained pattern. During the retrieval phase, presentation of just part of the input pattern will lead, via the auto-associative process just described, to more and more of the learned pattern. The process by which the network settles into a state corresponding to this previously learned pattern can be understood as involving movement into an attractor state in an attractive landscape, the shape of which is specified by the dynamical equations describing the operation of the network. Networks of this sort have been used to model a number of psychological or neurobiological processes including the recall of complete memories from partial cues (See, e.g. Trappenberg, 2002). Processing of this kind is often associated with brain structures such as the hippocampus. Such models obviously abstract away from many neural details, and in this respect are relatively non-mechanistic in Craver's sense.[19] On my view, however, we should not conclude that they are

[19] To the extent that such models explain in terms of generic facts about the structure of attractive landscapes and so on, they also involve abstraction away from the details of individual trajectories taken by the system in reaching some final state. That is, the explanation for why the system ends up in some final state has to do with, e.g. this being in a basin of attraction for the landscape, with the details of the exact process by which

unexplanatory for this reason alone. Instead their explanatory status depends on whether they accurately capture the dependency relations in real neural structures. This depends in turn on whether the modeled neural structures have the connectivity of a recurrent network, whether they involve Hebbian associative learning, whether there is empirical support for separate training and retrieval phases, and so on[20].

## 8. Mechanism, Predictivism and Instrumentalism

So far I have not addressed an important set of objections, due to Craver and others, to the ideas just defended. These objections turn on the claim that if we abandon the idea that explanation (at least in neuroscience) must be mechanistic, we lose the ability to make various important distinctions. For example, we lose the distinction between, on the one hand, purely descriptive or phenomenological models, and, on the other hand, explanatory models. We also lose the related distinction between the use of models, construed instrumentally merely for predictive purposes, and their use under realistic construals to explain. Craver argues, for example, that without the mechanistic constraints on explanation that he favors, we will be forced to regard Ptolemaic astronomy or models that merely postulate correlations as explanatory. Although it should be obvious from my discussion above that I disagree with many of these claims, I also think that they raise many interesting issues that are especially in need of discussion with the interventionist framework, since they often turn on what can be a possible target of intervention, when a model can be thought of as telling us what would happen under interventions, and when a model provides information about dependency relations in the relevant sense. In what follows I explore some of the different ways in which, from an interventionist perspective, a model may be merely descriptive or phenomenological rather than explanatory. This will give us a sort of catalog of different ways in which models can be explanatorily deficient, but, as we shall also see, a model can avoid these deficiencies without being mechanical.

(1) Obviously one straightforward way in which the interventionist requirements can be violated is that the factors cited in some candidate explanans correspond to "real" features *F* in the world, but the model should not be understood as even attempting to describe how explanandum *E* responds to interventions on those features or as describing a dependency relation (in the relevant sense) between *F* and *E*. This will be the case, for example, for models in which the relationship between *F* and *E* is (and is understood to be) purely correlational rather than causal. For example, a model might represent the correlation between barometer readings *B* and the occurrence *S* of a storm, and this representation may be descriptively accurate and predictively useful even though the *B*- *S* relationship is not causal. The non-causal, non-explanatory status of such a model follows, within the interventionist framework, from the fact that the model does not tell us how (or even whether) *S* will change under interventions on *B* or about a dependency

the system falls into that state being omitted from the model. This is arguably another respect in which the system departs from some of the expectations we have about mechanical explanations, since specific trajectories are often taken to matter for these.

[20] For additional relevant discussion concerning a different neural network model (the Zipser- Andersen Gain Field model) see Kaplan, 2011,Section 7.

relation between *B* and *S*. Note that reaching this judgment does *not* require acceptance of the idea that only models that are mechanistic in the sense of section 7 above or that provide lots of implementational detail explain: a model that abstracts away from such detail can nonetheless describe relationships that are causal in the interventionist sense (or that are explanatory in the sense of describing dependency relationships) and a purely correlational model might include lots of detail about the material composition of the modeled system and the spatio-temporal organization of its parts.

(2) A different, and in some respects more interesting, kind of case arises when a theory or model is interpreted as (or purports to) describe dependency relationships that but these completely fail to track the actual dependency relations operative the system whose behavior the theory purports to explain. Of course models of this sort can nonetheless be descriptively accurate and predictively useful to the extent that they correctly represent correlational patterns among variables

A plausible example of this possibility, discussed by Kaplan and Craver (2011) is Ptolemaic astronomy. According to this theory (at least in the cartoon version we consider here) the planets move as they do because they are carried around in their orbits by revolving crystalline spheres centered on the earth, or by additional crystalline spheres ("epicycles") whose centers move on the geocentric revolving spheres. It is uncontroversial that nothing like such spheres exists and that the motions of the planets do not depend on their being carried around on such spheres. There is thus no legitimate interventionist interpretation of Ptolemaic astronomy as correctly telling us what would happen to the planetary orbits if interventions were to occur on such spheres (e.g., by changing their rates of revolution or disrupting them in some way.) Nor does this theory provide other sorts of explanatorily relevant information about dependency relationships. not exist[21]. It follows that Ptolemaic astronomy does not qualify as an explanatory theory within the interventionist framework. It is a purely phenomenological (or descriptive) theory, although for somewhat different reasons than the barometer reading/storm "theory" discussed under section 1 above.

The case of Ptolemaic astronomy seems clear enough but there are many other examples involving models with "unrealistic" elements that raise subtle and interesting questions regarding their explanatory status. Although I lack the space for detailed discussion, my general view is that a model can contain many features that do not directly correspond to or mirror features of a target system but nonetheless be explanatory in virtue of correctly characterizing dependency relations governing that system. On my

[21] I would thus reject (as a general condition on explanation) condition (a) in Kaplan and Craver's 3M requirement, which holds in that "[i]n 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…" (p. 611). I am much more sympathetic to their second condition (b), when properly interpreted: "(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". I will add, though, that condition (a) may have more plausibility when construed more narrowly as a requirement on what it means for an explanation to be "mechanistic".

view, what matters most for purposes of explanation is that the model correctly characterizes dependency relations relevant to the explananda we are trying to explain. That the model may misrepresent *other* dependency relations relevant to *other* explananda that the model does not attempt to explain or that it mischaracterizes in some respects (or cannot be taken literally in what it says regarding) the entities or properties standing in those relations often matters less much from the point of view of explanation. To take a simple example, a network model in which neurons are represented as interconnected dimensionless points may nonetheless correctly describe what would happen to the network or how it would behave under various changes in the inputs delivered to those neurons (so that the model is explanatory with respect to these explananda), even though it is of course true that neurons are not dimensionless points and some predictions based on this assumption will be obviously mistaken. As another illustration, it is arguable Bohr's model of the atom had some explanatory force in virtue of correctly representing the dependency of the emission spectrum for hydrogen on transitions between electronic energy levels (and the dependency of the latter on the absorption of photons), even though in other respects the model was representationally quite inaccurate. For this reason, I do not think that it is correct to claim that if model is to provide satisfactory explanations all of the variables in the model must correspond directly to entities or properties that are present in the target system[22]. Models can successfully convey dependency information in surprisingly indirect ways that do not require this sort of mirroring or correspondence of individual elements in the model to elements in the world. I acknowledge that this introduces a certain vagueness or indeterminacy into assessments of explanatory status (when is a model so far "off" in what it claims about the target system that we should regard it as unexplanatory) but I believe this to be unavoidable.

3) Yet another possibility is that a theory or model might be merely descriptive in the sense that it describes or summarizes a pattern in some body of data in terms of variables *X*, *Y* etc, but without any suggestion that these variables are related causally in the interventionist sense. For example, a model according to which the distribution of velocities of molecules in a gas is Gaussian is merely descriptive in this sense, as is a model according to which the receptive fields of neurons can be represented by the difference between two Gaussians—an example considered in Kaplan and Craver (2011). A closely related possibility is that the model simply describes some regularly occurring phenomenon but without telling us anything about the factors on which the occurrence of that phenomenon depends, as was the case for the "phenomenological" representation of neural tuning curves discussed in section 2.

[22] To put the point in a slightly different way, whether a model gets the underlying ontology of the target system right and whether it conveys correct information abut dependency relations and the answers to what-if things- had been –different questions are much more independent of one another than many philosophers suppose. On my view, it is the latter (getting the appropriate relationships rather than the relata) that matter for explanation. A version of the wave theory of light that conveys correct information about relationships (including intervention supporting relationships) involved in reflection, refraction, diffraction and so on should be regarded as explanatory even if the theory represents waves themselves as mechanical displacements in an ether.

(4) The model might describe a predictively useful relationship which involves one or more variables that are not, for logical or conceptual reasons, possible targets for intervention. An illustration (due to Kaplan 2011) is provided by the Balmer formula which gives the wavelength ( $\lambda$) of lines in the absorption/emission spectrum of hydrogen in terms of the relation: $\lambda = B\ (m^2/m^2-4)$ where $B$ is a constant and $m$ an integer greater than two. This relationship is not a causal relationship, at least according to the interventionist account, since the notion of intervening to change the value of $m$ from one integral value to another does not make sense. We cannot interpret the Balmer formula as telling us what would happen to $\lambda$ under interventions on the number $m$. Nor does this seem to be a case of a dependency relationship of any other kind relevant to explanation.

(5) Another possible way in which the interventionist requirements can fail is that a theory or model can be so unclear or non-committal about how some of the terms or variables in the theory are to be interpreted (or what features they correspond to in the world) that we have no conception of what would constitute an intervention on those features, what would happen under such an intervention, or even what would be involved in those features varying or being different. (This possibility contrasts with the case of Ptolemaic astronomy described under 2) since it seems clear in a general way what crystalline spheres would be were they to exist, and what would be involved in their varying in diameter and position and so on.) An extreme case is a theory which is just a mathematical structure or an entirely uninterpreted set of equations relating certain variables. To the extent that the theory does not specify at al what structures or relations in the world are supposed to correspond to the dependency relationships postulated in the theory, then, according to the interventionist framework, it is not even a candidate for an explanatory theory. (For example, the HH model, considered simply as a set of equations without any physical interpretation, is not even a candidate for an explanation.) Another, less extreme possibility along these lines is that the theory does not contain completely uninterpreted variables and relationships but instead provides some characterization of these, perhaps giving them a semantic label or even assigning a number to them, estimated from other measured quantities, but nonetheless leaves their physical or wordly interpretation sufficiently underspecified that we lack any clear conception of what would be involved in intervening on them or what corresponds in the target system to the dependency relations in which they figure. The "gating' variables fitted by Hodgkin and Huxley to the expressions describing the voltage and time dependencies of sodium and potassium channels in their model of the generation of the action potential had something of this character, as discussed below (Section 9) .

Another related possibility is represented by the treatment of bimanual coordination by Haken et al. 1985 (the HKB model), which is championed by Chemero and Silberstein, 2008 as an alternative to more standard mechanistic or computational accounts of psychological and neuroscience explanation. When subjects attempt to move their left and right index fingers in phase in time with a metronome, their movements are found to be related by

$$(8.1)\ d\O/dt= - a \sin\O -2b \sin 2\ \O$$

where $\O$ is the relative phase angle between the two fingers and $b/a$ reflects the finger oscillating frequencies. It is readily seen that this equation permits just two stable outcomes, when either $Ø$ = 0 or $Ø$= 180 degrees, corresponding to the movement of fingers either in-phase (parallel, like windshield wipers) or in anti-phase. As *b/a* decreases (corresponding to faster finger oscillation), subjects are unable to maintain the antiphase movement and switch to the in-phase movement, with this being regarded as a "phase transition". This behavior is reflected in the basins of attraction associated with (8.1); there are two attractors (at *Ø = 0* or *Ø= 180*) when *b/a* is relatively large and just one when this ratio is small.

I agree with Kaplan and Craver (2011) that it is difficult to see this as a causal or as an explanatory model[23]. To begin with, it does not purport to tell us anything about the neural features on which the described behavior depends—in this respect, it seems like a non-starter as an example of neuroscientific or psychological explanation and, contrary to what Chemero and Silberstein claim, a dubious candidate for a replacement for such explanations. Because there is no accompanying neural account (indeed, as far as the model itself goes, no claim about whether such an account even exists), it is unclear how, if at all, to interpret the HKB model as a causal or explanatory model. As far as the model and the accompanying experimental data go, the restricted possible states of coupled finger movement and the "phase transition" might be due to some common neural/nervous system cause, in which case these aspects of the phenomenon will have more of the character of a correlation among joint effects than a causal relationship. Indeed, Kelso himself in his 1984 paper proposes that the relation (8.1) may be regarded as "constrain[ing] possible neural explanations" (p. 93) of the facts about finger movement he describes, which suggests that (8.1) has more of the status of a potential explanandum for a genuinely explanatory theory (or an empirical constraint on such a theory) grounded in more general features of the brain or nervous system, rather than something which should itself be regarded as explanatory[24].

The cases 1-5 are all cases in which the interventionist requirements for explanation are not met. Note, however, that none are cases in which a theory or model fails to be explanatory simply because it fails to provide extensive mechanistic or implementational detail. Instead, at least from an interventionist perspective, the models under 1-5 fail to be explanatory for other, independent reasons – because they invoke merely correlational relationships or non-existent or woefully underspecified dependence

[23] Although I do not regard the HKB model as a plausible example of an explanatory psychological/neuroscientific model rooted in dynamic systems theory, I emphasize, as argued above, that in my view it would be a mistake to suppose that all dynamic systems accounts of brain function in terms of attractor landscapes and the like are non-explanatory. In addition to the theories of memory retrieval mentioned above, other plausible candidates for explanatory models involving dynamic systems theory include accounts of categorization and decision-making of the sort described in Rolls and Deco, 2010.

[24] I will also add that the motivation for (1) in Haken et al's (1985) paper also does not seem to have much to do with distinctively causal considerations. Instead (8.1) is motivated by perceived "analogies" (rooted in "synergetics") with the behavior of other sorts of physical systems exhibiting phase transitions, with (1) described as the "simplest" equation (p. 47) of a certain general form subject to certain symmetry constraints that fits the observed data describing finger movements.

relations and so on. In other words, we can explain what is explanatorily defective about such models in terms of violations of basic interventionist/dependency requirements on explanation without invoking the idea that all explanations must be mechanistic. To the extent that a model avoids the problems described under 1-5 above, and satisfies the interventionist constraints on explanation, it will count as explanatory even if it fails to be mechanistic. For example, depending on the details of the case, a recurrent network model for auto-associative memory may describe genuine dependence relations in a target system (a brain) in the interventionist sense, rather than just correlations and the items related via these dependence relations—neurons, connections among neurons and neural activity – may be "real" and possible objects of intervention. It may also be clear enough what would be involved in intervening on such a structure (e.g. by changing its input or more dramatically by lesioning it) so the model is not one in which it is left completely unclear or unspecified what in the world corresponds to relevant variables. Similarly it may be clear enough what the relationships postulated in the model imply about what would happen in the target system under various manipulations or perturbations. On the other hand, the model lacks implementational or mechanistic detail, thus illustrating the independence of this feature from the kinds of deficiencies represented by 1-5.

## 9. The Hodgkin-Huxley Model

Many of the themes discussed above are illustrated by the Hodgkin-Huxley (hereafter HH) model, to which I now turn. This has been the subject of a considerable recent discussion, with some (e.g., Craver 2008 and Bogen 2008) regarding the model as unexplanatory (or in Craver's case, at best an explanation sketch) because of its failure to provide various sorts of mechanistic detail and others (Weber 2008, Levy, forthcoming) defending the explanatory status of the model. As will be seen, my own assessment is very close to that of Weber and Levy, and I will draw on both of their discussions in what follows.

I take the goal of HH's 1952 paper to be the presentation of a model of the generation of the action potential in an individual neuron. The experiments HH report were conducted on the giant axion of the squid, although it is assumed that many of the features of the model apply much more generally. The explanandum of the model is a phenomenon or stylized fact (in the sense described in section 3) having to do with shape of the action potential— what Trappenberg calls the "prototypical form of the action potential" (p. 33). This involves a change in the potential across the neuron's membrane which follows a characteristic pattern: first rising sharply to a positive value from the resting potential of the neuron (depolarization) and then decreasing sharply to below the resting potential, followed by a recovery to the resting potential. The action potential results from changes in the conductance of the membrane to sodium and potassium ions, with the rise in potential being due to opening of Na channels in the membrane leading to the influx in Na ions and the subsequent fall being due to the inactivation of the sodium channels approximately 1ms after their opening and the opening at this point of the potassium channels. These ionic currents are responsible for the patterns of change in membrane potential. Furthermore the channels themselves are "voltage-gated" with the channel resistances/ conductances being influenced by the membrane potential.

The basic idea of the H-H model is that structural features of the neuron responsible for the action potential may be represented by a circuit diagram with the following structure:

Figure 1 here

This is a circuit in parallel with (reading from left to right) a capacitor which stores charge (the potential across the membrane functions as a capacitor), a channel[25] that conducts the sodium current $I_{Na}$, with an associated time and voltage dependent conductance $g_{Na}$, a channel that conducts a potassium current $I_K$ with time and voltage dependent conductance $g_K$, and a leakage current $I_l$ which is assumed to be time and voltage independent. The relationships governing these quantities are represented by HH by means of a set of differential equations. First, the total membrane current $I$ is written as the sum of the capacitor current and the total ionic current $I_i$:

$I= C_m dV/dT+ I_i$ (This is just a version of Kirchoff's law for the conservation of charge.)

The ionic current in turn is the sum $I_i = I_{Na} + I_K +I_l$

These last three currents can be written as $I_{Na} = g_{Na} (V-V_{Na})$ , $I_K = g_K (V-V_K)$ , and $I_l =g_l (V-V_l)$ where $V_{Na}, V_k, V_l$ are the equilibrium membrane potentials. These are just versions of Ohm's law, with the currents being equal to the products of the conductances and the difference between the membrane potential and the equilibrium potential. The ionic conductances in turn are expressed as the product of the maximum conductances (which I will write as $G^*_{Na}$ etc. for the channels) times "gating" variables $n$, $m$, and $h$:

$G_k = G^*_K n^4$
$G_{Na}= G^*_{Na} m^3 h$

The underlying picture is that the passage of ions through a channel requires the opening of a number of distinct hypothetical structures or "gates", with the gating variables representing the probability that these are open. For example, $n$ represents the probability that a gate in the potassium channel is open, it is assumed that four distinct gates must be open for the passage of the potassium current, and also that these gates open independently, so that $n^4$ is in effect the probability that the potassium channel is open. $G^*_K n^4$ thus yields an expression for the active or available conductive as a function of the maximum conductance. Variables $m$ and $h$ have similar interpretations: the Na current requires that three gates, each with probability $m$, be open and that a distinct gate also be open with probability $h$. Other equations, not reproduced here, describe the time

[25] As noted above, the channels which these variables in the H-H model describe are really (from a molecular perspective) aggregates or classes of channels of various types (Na etc.) rather than individual ion currents.

derivatives of the gating variables *n* etc. as functions of other variables such as the voltage dependent opening and closing rates of the gates.

Combining these equations yields:

$$(9.1)\ I = C_M dV/dt + G^*_K n^4(V - V_K) + G^*_{Na} m^3 h(V - V_{Na}) + G_l(V - V_l)$$

$G^*_{Na}$ and $G^*_K$ are directly measured variables but, by HH's own account, the gating variables (and the variables occurring in the differential equations describing how these change with time) were chosen on the basis that they fit the experimental data reasonably well and were simple. Lacking information about the details of the molecular mechanisms governing the operation of the channels, HH in effect settled for expressions (the quantities *m*, *n* and *h*, the powers to which these are raised, and the equations specifying the time course of these) that accurately empirically described the channel conductances, and, although they speculated on possible physical interpretations for these expressions, they did not claim that they had successfully identified the mechanisms responsible for them. They write " the success of the equations[26] is no evidence in favor of the mechanism of permeability changes [i.e. changes in membrane conductance] that we tentatively had in mind when formulating them" (p. 541). On the other hand, the passage just quoted is immediately followed by this remark (also quoted by Weber and by Levy):

> The point that we do consider to be established is that fairly simple permeability changes in response to alterations in membrane potential, of the kind deduced from the voltage clamp results, are a sufficient explanation of the wide range of phenomena that have been fitted by solutions of the equations. (p. 541)

Indeed, their entire 1952 paper is full of language strongly suggesting that they think of themselves as having provided a causal explanation or a causal account of the action potential. Their introductory paragraph says that their model "will account for conductance and excitation in quantitative terms" (p. 500) and the first page of their paper contains language like the following:

> Each component of the ionic current is *determined by a driving force* which may conveniently be measured as an electrical potential difference and a permeability coefficient. (p.500, emphasis added)
>
> The *influence* of membrane potential on permeability can be summarized by stating: first, that depolarization *causes* a transient increase in sodium conductance and a slower but maintained increase in potassium conductance;

[26] I follow Weber in interpreting the reference to "the equations" in this passage to the equations HH propose describing the dependence of the channel conductances on *m*, *n*, and *h* and to the equations describing the time dependence of the latter, rather than to the equation (9.1)

> secondly, that these changes are graded and that they can be *reversed* by repolarizing the membrane. (p. 500, emphasis added)

They go on to say that:

> In order to decide whether these effects *are sufficient to account* for complicated phenomena such as the action potential and refractory period, it is necessary to obtain expressions relating the sodium and potassium conductances to time and membrane potential (page 500-1, emphasis added)

The judgment that the HH model is explanatory is repeated in many if not most of the papers and texts I consulted that contain explications of the model. For example, in the passage quoted from Herz et al. above (Section 6), the HH model is described as "explaining" and providing "quantitative understanding". McCormack (2003) writes that the experiments and model in the 1952 paper "explained qualitatively and quantitatively the ionic mechanism by which the action potential is generated" (p. 145). Koch (1999) writes that "the biophysical mechanisms and underlying action potential generation in the cell body of both vertebrates and invertebrates can be understood and modeled by the formalism Hodgkin and Huxley introduced.." (p. 144).[27] Similarly, Trappenberg (2002, pp 34ff) repeatedly characterizes the HH model as describing the "mechanism" (or "minimal mechanism') for the generation of the action potential.

I follow both Weber and Levy in holding that the obvious way of reconciling HH's various remarks about the explanatory status of their model is to distinguish the question of whether HH provided (i) an explanation of the generation of the action potential from the issue of whether they provided (ii) a satisfactory explanation of the operation of the ion channels and the molecular mechanisms involved in gating. Both by their own account and judged in the light of subsequent understanding of the operation of the ion channels, they do not provide (ii). However, as argued in previous sections, this is consistent with their having provided an explanation of (i) the generation of the action potential. Put at a very general level, this is because the equation (9.1) and the associated model identifies the factors (or at least many of the factors) on which the generation of the action potential depends, although it does not successfully identify (or at least very fully or adequately identify) the factors on which the operation of the ion channels depends. The possibility of explaining (i) without explaining (ii) can be thought of as reflection of the general point, made in previous sections in connection with modeling strategies, that models work at different levels or scales, and a model can explain some explananda at a particular scale or level (the overall behavior of behavior of a neuron in

[27] I should also acknowledge, though, that this remark by Koch is followed shortly by a reference to the "phenomenological model.. of the events underlying the generation of the action potential" (144-5) postulated by HH, which seems to mix together the claim that the model provides causal information ("generation") with a description of it as "phenomenological". This makes sense if "phenomenological" in this context just means "lacking lower level mechanistic detail" (which is not taken to imply that the account is non-causal or non-explanatory). This is perhaps the sense in which classical thermodynamics is a "phenomenological" theory.

generating an action potential) without explaining aspects of neural behavior at other scales or levels (the molecular mechanisms associated with the ion channels).

As Trappenberg suggests, one way of thinking of the HH model is as a kind of minimal model of the generation of the action potential. The HH model shows that the generation of the action potential depends on (or requires at a minimum), among other things, the existence of at least two voltage gated and time-dependent ion channels, as well as an additional static or leakage channel and a membrane that is otherwise sufficiently insulated to act as a capacitor. However, given that such a structure is present and behaves appropriately, the presence of the specific mechanism by which the ion channels in the giant squid operates is not required for the generation of the action potential, as long as some mechanism or other that plays this role is present. This in effect allows for the separation of explanatory tasks (i) and (ii) in the manner that I have described.

This assessment of the explanatory status of the HH model also follows from the interventionist requirements on explanation described in section 2 – a point that is also developed by Weber (2008). For example, the HH model correctly describes what will happen to the total current $I$ under interventions on the transmembrane voltage $V$ (which can be accomplished experimentally via the voltage clamp device), and under changes in the maximum sodium and potassium channel conductances, which can be accomplished by techniques for molecular manipulation of these. Although the HH model does not correctly describe the molecular mechanisms involved in the operation of ion channels, it does claim, correctly, that it should be possible to intervene on these classes of channels independently and to change the individual currents, $I_{\mathrm{Na}}$ and $I_{\mathrm{K}}$, independently of each other and independently of the other terms in equation. The equation and associated correctly describes what would happen to the total current under such interventions. The HH model is thus (at least in this respect) modular and effects a decomposition of the structure responsible for the membrane current into components, each of which is governed by generalizations which operate independently of the generalizations governing the other components. In this sense it seems fairly natural to characterize the HH model as describing the "mechanism" of the action potential, as a number of the writers quoted above do.

We may also note that, putting aside the role of the gating terms and the equations governing them, the HH model does not exhibit any of the pathologies described in section 8 which render a model merely descriptive or phenomenological rather than explanatory. In particular, the HH model does not (i) describe a relationship (between $I$ and terms like $V, I_{Na},..)$ that is purely correlational rather than causal in the interventionist sense. Moreover, with the partial exception of the gating terms, the relations among other terms conveys information about dependency relations in the target system. For instance, $V$, the various currents, the membrane capacitance, and the sodium and potassium conductances all refer to features of the world that are "real" in the sense that they can be measured and manipulated and the model correctly describes how these features are related (via intervention-supporting dependency relations) to one another in the target system. In these respects, the HH model is very different from the Ptolemaic model.

## 10. Conclusion.

In this paper I have attempted to use an interventionist framework to argue that theories and models in neurobiology that abstract away from lower level or implementational detail can nonetheless be explanatory. I have tried to show that this conclusion does not require that one abandon the distinction between models that are explanatory and those that are merely descriptive or predictively accurate, but non-explanatory. Instead interventionism provides a natural framework for capturing this distinction. I have also argued that mechanistic models are just one possible form of explanatory model; they are explanations that meet certain additional conditions that qualify them as "mechanistic". Models that are not mechanistic can nonetheless count as explanatory if they correctly capture dependency relations that support interventions.

**References**

Batterman, R. (2009) "Idealization and Modeling," *Synthese, 169*, 427-446.

Batterman, R. and Rice, C. (2014) "Minimal Model Explanation" *Philosophy of Science* 81: 349-376.

Bechtel, W. and Richardson, R, *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research.* Princeton: Princeton University Press.

Bechtel, W. and Abrahamsen, A. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, 18, 707-723.

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

Bogen, J. (2005) 'Regularities and Causality; Generalizations and Causal Explanations', *Studies in History and Philosophy of Biological and Biomedical Sciences* 36, pp. 397–420.

Bogen, J. (2008) "The hodgkin- huxley equations and the concrete model: Comments on Craver, Schaffner, and Weber" *Philosophy of Science* 75 (5):1034-1046.

Chemero, A, and Silberstein. M. 2008. "After the Philosophy of Mind: Replacing Scholasticism with Science." *Philosophy of Science* 75:1–27

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

Craver, C. F. [2006]: "When Mechanistic Models Explain", *Synthese,* 153: 355-376.

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

Dayan, P. and Abbott, L. (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural System*s. Cambridge, MA: MIT Press.

Haken, H., Kelso, J. and Bunz, H. (1985) " A Theoretical Model of Phase Transitions in Human Hand Movements" *Biological Cybernetics* 51: 347-442.

Herz, A. Gollisch, T . Machens, C. Jaeger, D. (2006) "Modeling Single-Neuron Dynamics and Computation: a Balance of Detail and Abstraction" *Science* **314**, 80- 85.

Hodgkin, A. and Huxley, A. (1952) "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve" *Journal of Physiology* 117: 500-544.

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

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

Levy, A. (Forthcoming) "What was Hodgkin and Huxley's Achievement?" *British Journal for the Philosophy of Science*

Levy, A. (forthcoming) "Causal Organization and Strategies of Abstraction"

Levy, A. and Bechtel, B. (Forthcoming) "Abstraction and the Organization of Mechanisms" *Philosophy of Science*.

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

McCormack, D. "Membrane Potential and Action Potential" in Squire, L., Bloom, F. McConnell, S. Roberts, J. Spitzer, N. and Zigmond, M. (2003) *Fundamental Neuroscienc*e. San Diego: Academic Press.

Piccinini, G. and Craver, C. 2011 " Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches" *Synthese* 183 :283-311.

Rolls, E and Deco, G. (2010) *The Noisy Brain: Stochastic Dynamics as a Principle of Brain Functioning*. Oxford: Oxford University Press.

Sternberg, S. (2001) "Separate Modifiability, Mental Modules, and the Use of Pure and Composite Measures to Reveal Them" *Acta Psychologica* 106: 147-246.

Strevens, M. (2008) Depth: *An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Thomson-Jones, M. (2005). "Idealization and Abstraction: A Framework." In M. Thomson-Jones and N. Cartwright (eds.), *Idealization XII: Correcting the Model*. Amsterdam: Rodopi, pp. 173-217.

Trappenberg, T. (2002) *Fundamentals of Computational Neuroscience*. Oxford: Oxford University Press.

Weber, M. (2008): “Causes Without Mechanisms: Experimental Regularities, Physical Laws, and Neuroscientific Explanation” *Philosophy of Science* 75: 995-1007.

Woodward, J. (1979) “Scientific Explanation” *British Journal for the Philosophy of Science* 30: 41-67.

Woodward, J. (2002) “What is a Mechanism? A Counterfactual Account” *Philosophy of Science*, 69: S366–S377.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press.

Woodward, J. “Comments on John Campbell’s Causation in Psychiatry in Kendler and Parnas, (eds.) *Philosophical Issues in Psychiatry: Explanation, Phenomenology and Nosology* Johns Hopkins University Press, 2008, pp 216- 235.

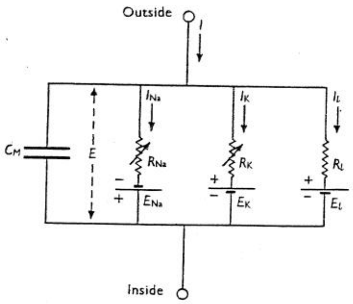

Figure 1