ORIGINAL PAPER

When mechanistic models explain

Carl F. Craver

Abstract   Not all models are explanatory. Some models are data summaries. Some models sketch explanations but leave crucial details unspecified or hidden behind filler terms. Some models are used to conjecture a how-possibly explanation without regard to whether it is a how-actually explanation. I use the Hodgkin and Huxley model of the action potential to illustrate these ways that models can be useful without explaining. I then use the subsequent development of the explanation of the action potential to show what is required of an adequate mechanistic model. Mechanistic models are explanatory.

Keywords   Mechanisms · Explanation · Models · Electrophysiology · Action · Potential · Hodgkin · Huxley · Functional Analysis

1 Introduction

There is a widely accepted distinction between merely *modeling* a mechanism's behavior and *explaining* it. Ptolemy's models of planetary motion predict the paths of the planets through the night sky, but they do not explain their trajectories. Tinker-toy models simulate human tic-tac-toe competence, but nobody believes that a blueprint of the model's grinding gears explains how human beings play. Models play many roles in science beyond providing explanations (Bogen, 2005). They are used to make precise and accurate predictions. They are used to summarize data. They are used as heuristics for designing experiments. They are used to demonstrate surprising and counterintuitive consequences of particular forms of systematic organization. But some models have an additional property beyond these others: they are explanations.

C. F. Craver (✉)
Department of Philosophy and Philosophy–Neuroscience-Psychology Program,
Washington University, 1 Brockings Drive,
St. Louis, MO 65130, USA
e-mail: ccraver@artsci.wustl.edu

My goal is to articulate some of the ways that models, even very useful models, can fail to provide explanations. I also provide a general account of when mechanistic models explain. My account is driven by a curious fact about the history of electrophysiology. In 1952, Hodgkin and Huxley published a mathematical model of the action potential in the squid giant axon. The model is derived in part from laws of physics and chemistry, such as Ohm's law, Coulomb's law, and the Nernst equation, and it can be used to derive myriad electrical features of many different kinds of neurons in many different species. Despite this accomplishment, the authors insist that their model is not an explanation. This is curious if one thinks, as many did at the time, that to explain a phenomenon is just to show that it follows from laws of nature and specification of the antecedent and background conditions. I argue that Hodgkin and Huxley regarded their mathematical model as a phenomenological model and that they regarded their understanding of the action potential as sketchy at best. This phenomenological sketch was gradually transformed into an explanation as researchers discovered the details of the underlying mechanism. The Hodgkin and Huxley model is a potential historical challenge to recent mechanistic views of explanation in biology and neuroscience. At first glance, it appears to be an exemplar of the covering-law model of explanation. I show that this appearance is misleading. Before discussing Hodgkin and Huxley's model, I first lay out three fundamental, and largely uncontroversial, distinctions: (1) between phenomenal models and explanations, (2) between sketches and complete explanations, and (3) between how-possibly models and how-actually models. In the final section of the paper, I say what is required to move beyond a mere model and to provide an adequate mechanistic explanation.[1]

2 Merely phenomenal models versus explanations

There are many senses of the term "model" in science, and these different kinds of model serve distinct scientific ends (e.g., Giere, 1999; Morgan & Morrison, 1999; Suppe, 1989, Suppes, 1967). My focus is on representational models, that is, those that scientists construct as more or less abstract descriptions of a real system. The following skeletal account of representational models is intended to be broad enough to include, for example, Ptolemy's model of the heavens, Bohr's model of the atom, Harvey's model of the circulatory system, Boyle's model of ideal gases, Watson and Crick's helical model of DNA, Ferree and Lockery's (1999) model of chemotaxis in *C. elegans*, and Rumelhart and McClelland's (1986) model of the generation of past-tense verbs.

The skeletal account is as follows. Take some feature (T) of a target system. T might be a static property of the system or it might be characterized as a mapping from inputs (or sets of inputs) onto outputs (or sets of outputs) implemented by a system. Inputs may include any condition judged to be relevant to the purposes of the model, including triggering conditions (such as the presentation of a present-tense verb) and background conditions (such as priming, attentional load, fatigue). Likewise, the outputs may include a host of outputs from the target output (e.g., the past-tense verb) and byproducts of the system's operation (e.g., glucose utilization or

---

[1] I will focus exclusively on constitutive explanation, in which a property or behavior of a system is explained by the properties and activities of its parts. I am especially concerned with mechanistic explanations, one subclass of constitutive explanations.

dissipated heat). Different purposes lead the modelers to include different inputs and outputs.

Modeling T involves constructing an algorithm or function (S) that generates a mapping from inputs onto outputs that is reasonably similar to T. The algorithms or procedures might be implemented in physical systems, written in computer programs, captured in mathematical equations, or sketched in block and arrow diagrams. All that matters is that (i) for each input (or set of inputs) in T there is a corresponding input (or set of inputs) in S, (ii) for each output (or set of outputs) in T, there is a corresponding output (or set of outputs) in S, and (iii) for each input–output relation in T there is a corresponding input–output relation in S.

T and S may be more or less similar, depending on one's skill at modeling and one's reasons for building the model. For some purposes, all one requires of a model is that it be *phenomenally adequate*. That is, the input–output mapping in S should be sufficiently similar to the input–output mapping in T for one's needs. Few models are actually isomorphic with the phenomenon, given that models typically abstract away from the precise details of the system being modeled, that they typically are only approximate, and that they make simplifying assumptions in order to apply a particular formalism. The weaker standard that the input–output mapping in T should be homomorphic with the mapping in S can be easier or harder to satisfy depending on how much detail one includes about the target phenomenon and on how similar one expects the model and the phenomenon to be. (See Piccinini (forthcoming) for a detailed discussion of why models are typically approximate.) The richer and more fine-grained one's characterization of the target phenomenon, the more the space of possible models for the phenomenon is constrained, and so the more challenging it is to build a phenomenally adequate model.

One might, for example, build a model that is useful only within a narrow range of conditions (such as health, proper functioning, or the absence of disturbing outside forces) but that fails outside of those narrow conditions. For example, one might provide a model of verb-tense generation that performs perfectly well when the brain and vocal cords are working properly, but that provides no insight into how the system will behave if something breaks or if the system is in extreme environmental conditions. Newton's laws of motion work well enough for many standard purposes but they fail to deliver appropriate results at speeds approaching the speed of light. Churchland (1989) similarly claims that sentential models of the mind fail to explain such paradigmatically mental phenomena as mental illness, emotion, learning, and phantom limbs. As another example, early advocates of connectionist simulations stressed that their models run quickly, operate on degraded inputs, and are robust in the face of damage, and so account for a host of phenomena that classical artificial intelligence models do not even pretend to encompass (McClelland & Rumelhart, 1986). The assumption behind such arguments is that a phenomenally (and so explanatorily) adequate model must account for all aspects of the phenomenon, not merely part of it. One sign of a *mere* model (i.e., one that is not explanatory) is that it is phenomenally adequate only for a narrow range of features of the target phenomenon. More accurately, however, phenomenal adequacy and explanatory force can and do vary independently of one another.

A model can be richly phenomenally adequate and non-explanatory. This is the take-home lesson of the several decades of attack on covering-law models of explanation at the hands of advocates of causal–mechanical models of explanation: merely subsuming a phenomenon under a set of generalizations or an abstract model does not

suffice to explain it. One can reliably predict the sunrise with a model of the circadian rhythms of a rooster, but the behavior of roosters does not explain the sunrise. One can predict sunny weather by looking at a barometer, but the behavior of the barometer does not explain the sunny weather. Similarly, one can build an unbeatable tinker-toy model of tic-tac-toe that simulates competent tic-tac-toe players, but the model does not explain human tic-tac-toe performance.

One need not resort to artificial examples. Ptolemaic models of the solar system allow one to predict the location of planets in the night sky, but nobody believes that the elaborate system of epicycles, deferents, equants, and eccentrics explains why the planets move. One can use Balmer's formula[2] to calculate the wavelengths of the emission spectrum for hydrogen, but Balmer's formula does not explain why the spectral lines show up where they do (see Cummins, 2000; Hempel, 1965). One can know how to apply Snell's law to predict how a beam of light will bend when it passes through a piece of glass without understanding anything about why light bends when it passes from one medium to the next. In Sect. 5, I show that Hodgkin and Huxley, and other electrophysiologists at the time, thought that their model of the action potential failed as an explanation in roughly the way that Ptolemy's models and Balmer's formula do, and I use the example of Hodgkin and Huxley's model below to get a better picture of how explanations and phenomenal models differ.

Such examples demonstrate that some phenomenal models appear to be non-explanatory and, further, that they have been treated as such in the history of science. But appearances and historical evidence cannot settle the normative question of whether phenomenal models ought to be considered explanatory. One might claim, after all, that Ptolemy's models *do* explain the motion of the planets, that Balmer's formula *does* explain the emission wavelength of hydrogen, and that Hodgkin and Huxley simply failed to recognize that their model explained the action potential. For someone who thinks this way, it will not help to multiply historical examples or to show that many contemporary scientists draw the same sharp distinction between explanatory models and models that are merely phenomenally adequate.

For now, I suggest an instrumentalist defense: Explanatory models are much more useful than merely phenomenal models for the purposes of control and manipulation. As Woodward (2003) argues, explanations afford the ability to say not merely how the system in fact behaves, but to say how it *will* behave under a variety of interventions (Woodward says to answer more "what-if-things-had-been-different" questions, or w-questions). Deeper explanations show how the system would behave under a wider range of interventions than do phenomenal models, and so they can be used to answer more w-questions. Because phenomenal models summarize the phenomenon to be explained, they typically allow one to answer some w-questions. But an explanation shows why the relations are as they are in the phenomenal model, and so reveals conditions under which those relations might change or fail to hold altogether. In that case, explanations outperform models that are merely phenomenally adequate because they cover a wider range of possible contingencies, afford a greater possibility of control over the phenomenon, and so allow one to answer a greater range of questions about how the phenomenon is dependent on various background conditions and underlying conditions.

---

[2] That is, $\lambda = 3645.6(n^2/(n^2 - 4))$. λ is wavelength, and n is any integer.

Indirect empirical evidence for the utility of moving beyond the phenomena to posit underlying explanations comes from Daniel Povinelli's (2000) contrast between causal reasoning in chimpanzees and human infants. Chimpanzees, Povinelli argues, are like Humeans. They do not posit hidden causal powers or mechanisms, and so they confine their understanding to regularities among the manifest events in their world. On relatively simple tasks, this strategy is highly effective. Chimpanzees are active manipulators of their environments, they mimic each other's behavior, and they quickly pick up on regularities between actions and consequences. Nonetheless, their understanding of even relatively simple systems never goes beyond the manifest sequence of events. Povinelli writes:

> . . . the range of concepts formed by the chimpanzee does not include concepts about entities or processes that have no perceptually based exemplars. On our view, chimpanzees detect the regularities that exist between events, and learn to act on the basis of them, but they do not appeal to unobservable phenomena (force, gravity, etc.) to account for (or assist in their reasoning about) such regular associations of events. Indeed, representations of hypothetical entities may be impossible without human-like language, or perhaps more directly, language may have created such representations. Thus, we accept the claim that chimpanzees form concepts about the world, but that their folk physics does not suffer (as Hume would have it) from an ascription of causal concepts to events which consistently covary with each other (298–299).

And it shows. Chimpanzees can be fooled, and can remain fooled for a very long time, by changes to a system that would never fool even a three-year-old. For example, chimpanzees can use a stick to dislodge a peanut from the inside of a clear plastic tube. When Povinelli changed the tube and placed the peanut to one side of a small trap in the tube's bottom, only three of seven chimpanzees learned to insert the stick from the side furthest from the peanut. Those that did solve the problem, however, seem to have relied on the crude rule that one should insert the stick on the side furthest from the reward. Flipping the tube, so that the trap is now on top and ineffective, failed to change the chimpanzee's behavior; they continue to insert the stick on the furthest side. The chimpanzees can see the spatial-layout of the tube, and they can learn that certain strategies do and do not yield peanuts, but they seem to be ignorant of the causal organization of the system. Povinelli describes similar failures across a spectrum of experimental tasks.

If Povinelli is right that Chimpanzees fail to form causal inferences, and if he is right about the impact of this failure on their ability to manipulate even simple devices, then his research displays poignantly the importance of moving beyond phenomenal models to models that describe underlying mechanisms. The chimpanzee's failure to reason about mechanisms leaves them unable to manipulate the system after even relatively minor changes. The chimpanzees thus point to a central contrast between *merely phenomenal models* and models that characterize the *mechanisms* responsible for the phenomenon. Ptolemy's models do not describe the mechanisms by which planets move. Balmer's formula does not describe the mechanisms by which the emission wavelength of hydrogen is what it is. Constitutive explanations go beyond merely describing the phenomenon. They describe the mechanism responsible for the phenomenon, that is, the mechanism that explains its diverse features (see Sect. 6 below).

3 Filler terms and sketches

Models that describe mechanisms can lie anywhere on a continuum between a
mechanism sketch and an *ideally complete description of the mechanism*.

A *mechanism sketch* is an incomplete model of a mechanism. It characterizes some
parts, activities, and features of the mechanism's organization, but it has gaps. Some-
times gaps are marked in visual diagrams by black boxes or question marks. More
problematically, sometimes they are masked by *filler terms*. Terms such as activate,
cause, encode, inhibit, produce, process, and represent are often used to indicate a
kind of activity in a mechanism without providing any detail about how that activity is
carried out. Black boxes, question marks, and acknowledged filler terms are innocu-
ous when they stand as place-holders for future work or when it is possible to replace
the filler term with some stock-in-trade property, entity, activity, or mechanism (as is
the case for "coding" in DNA).[3] In contrast, filler terms are barriers to progress when
they veil failures of understanding. If the term "encode" is used to stand for "some-
process-we-know-not-what," and if the provisional status of that term is forgotten,
then one has only an illusion of understanding. For this reason, neuroscientists often
denigrate the authors of such black-box models as "diagram makers" or "boxologists."

At the other end of the continuum are *ideally complete descriptions of a mecha-
nism*. Such models include all of the entities, properties, activities, and organizational
features that are relevant to every aspect of the phenomenon to be explained. Few
if any mechanistic models provide ideally complete description of a mechanism. In
fact, such descriptions would include so many potential factors that they would be
unwieldy for the purposes of prediction and control and utterly unilluminating to
human beings. Models frequently drop details that are irrelevant in the conditions
under which the model is to be used. Ideally complete mechanistic models are the
causal/mechanical analogue to Peter Railton's notion of an "ideal explanatory text,"
which includes all of the information relevant to the explanandum.

Which information is relevant varies from context to context. Explanations are
sometimes represented as answers to questions about why something has happened
or about how something works. Questions and answers, as linguistic entities, presup-
pose a conversational context that specifies precisely what is to be explained and how
much detail will suffice for a satisfying answer. This does not mean that the entire
search for an explanation is relative to a given pragmatic context: whether a given
piece of information conveyed in an answer is explanatorily relevant to the phenom-
enon specified in the conversational context is as much an objective fact about the
world as any other (more about this in Sect. 6). Which information is relevant varies
from context to context, but that a given bit of information is relevant in a particular
context is as objective a fact about the world as any other.

Between sketches and complete descriptions lies a continuum of *mechanism sche-
mata* that abstract away to a greater or lesser extent from the gory details (as Philip
Kitcher, 1984 calls them) of any particular mechanism. What counts as an appropriate
degree of abstraction depends, again, upon the uses to which the model is to be put.
Phenomenal models sit at the farmost limit of sketchiness. They are complete black
boxes; they reveal nothing about the underlying mechanisms and so merely "save the
phenomenon" to be explained.

---

[3] Stock-in-trade items (cf. Kauffman, 1971) are those that are accepted and understood by a science
at a time; they are included in the ontic store of the science (Craver & Darden, 2001).

4 From how-possibly to how-actually models

In order to explain a phenomenon, it is insufficient merely to characterize the phenomenon and to describe the behavior of some underlying mechanism. It is required, in addition, that the components described in the model should correspond to components in the mechanism in T.

Models vary considerably in their *mechanistic plausibility*. For those interested in building plausible simulations, it will not suffice for S simply to reproduce the input–output mapping of T. The model is further constrained by what is known about the internal machinery by which the inputs are transformed into outputs. It is possible, for example, to simulate human skills at multiplication with two sticks marked with logarithmic scales; but that is not how most humans multiply.

Early connectionists championed their models not just for their phenomenal adequacy but also on the basis of their biological plausibility in comparison to classical artificial intelligence (again see McClelland & Rumelhart, 1986). The claim was that the nodes in such networks are analogous to neurons, weight adjustments are roughly analogous to different forms of synaptic plasticity, and so on. Subsequent debate has revealed that connectionist models are themselves highly idealized and quite distant from the complex behavior of real neural networks, and neuroscientists have labored to build ever-more physiologically plausible models of the central nervous system. For those who merely want to predict the target system's performance, biologically implausible simulations will work just fine. But for those who build simulations in the search of explanations, mere simulation is not enough.

How-possibly models (unlike merely phenomenal models) are purported to explain, but they are only loosely constrained conjectures about the mechanism that produces the explanandum phenomenon. They describe how a set of parts and activities might be organized such that they exhibit the explanandum phenomenon. One can have no idea if the conjectured parts exist and, if they do, whether they can engage in the activities attributed to them in the model. Some computer models are purely how-possibly models. For example, one might simulate motion detection in LISP without any commitment to the idea that the brain is somehow executing CARs and CDRs (the basic operations of LISP). How-possibly models are often heuristically useful in constructing a space of possible mechanisms, but they are not adequate explanations. In saying this, I am saying not merely that the description must be true (or true enough) but further, that the model must correctly characterize the details of the mechanism in T. *How-actually models* describe real components, activities, and organizational features of the mechanism that in fact produces the phenomenon. They show how a mechanism works, not merely how it might work.

Between how-possibly and ideal explanations lies a range of *how-plausibly* models that are more or less consistent with the known constraints on the components, their activities, and their organization.[4] Again, how accurately a model must represent the details of the internal workings of a mechanism will depend upon the purposes for which the model is being deployed. If one is trying to explain the phenomenon, however, it will not do merely to describe some mechanisms that would produce the phenomenon. One wants the model, in addition, to show how T produces the phenomenon.

---

[4] The distinction between how-possibly, how-plausibly, and how-actually descriptions of mechanisms is introduced in Machamer, Darden, & Craver (2000).

Philosophers of the special sciences, such as Robert Cummins (1975, 1983), Daniel Dennett (1994), Bill Lycan (1999), Herbert Simon (1969), emphasize that explanations often proceed by functional analysis, reverse engineering, homuncular explanation, and decomposition. One begins with a complex phenomenon, and one shows how that phenomenon can be produced by teams of less capable sub-systems whose organized behavior composes the behavior of the system as a whole. The behavior of the sub-systems can often be explained in turn by postulating the behavior of various sub-sub-systems, and so on. As a first-pass description of the nature of explanation in sciences such as cognitive neuroscience, physiology, and molecular biology, this is a helpful descriptive framework. However, these accounts have yet to be developed to the point that they can distinguish how-possibly from how-actually functional analysis, reverse engineering, and homuncular explanation. Cummins (1975, 1983) sometimes speaks as if any description of a sequence of steps between input and output will suffice to explain a phenomenon. In speaking this way, Cummins erases the distinction between how-possibly and how-actually. Other times, he insists that such descriptions must ultimately bottom out in descriptions of neurological mechanisms (Cummins, 2000). According to the current view, constitutive explanations require descriptions of real mechanisms, not mere how-possibly posits.

I have now introduced three crucial distinctions for thinking about the relationship between models and explanations: the distinction between phenomenal models and explanations, the distinction between sketches and complete descriptions, and the distinction between how-possibly and how-actually explanations. In the next section, I show how the current understanding of the action potential developed as phenomenal models were elaborated, sketches were completed, and how-possibly explanations were jettisoned as part of the search for fundamental mechanisms in electrophysiology.

5 The Hodgkin and Huxley model

Hodgkin and Huxley's model of the action potential is now a cornerstone of electrophysiology and neuroscience. Action potentials are rapid and fleeting changes in the electrical potential difference across a neuron's membrane. This potential difference, known as the membrane potential (V_m), consists of a separation of charged ions on either side of the membrane. In the neuron's resting state, positive ions line up against the extracellular surface of the membrane, and negative ions line up on the intracellular surface. In typical cells, this arrangement establishes a polarized resting potential (V_{rest}) of -60 mV to -70 mV. In an action potential, the membrane becomes fleetingly permeable to sodium (Na^+) and potassium (K^+) ions. This allows the ions to diffuse rapidly across the cell membrane. This flux changes V_m. The action potential consists of a rapid rise in V_m to a maximum value of roughly $^+35$ mV, followed by a rapid decline in V_m to values below V_{rest}, and then an extended after-potential during which the neuron is less excitable (known as the refractory period).

Hodgkin and Huxley characterized the time-course of the action potential phenomenally in terms of the following features (modified from Hodgkin & Huxley, 1952, 542–543):

(a) the form, amplitude, and threshold of an action potential;
(b) the form, amplitude, and velocity of a propagated action potential;
(c) the form and amplitude of the resistance changes during an action potential;
(d) the total movement of ions during an action potential;
(e) the threshold and response during the refractory period;
(f) the existence and form of subthreshold responses;
(g) the production of action potentials after sustained current injection (that is, anodal break); and
(h) the subthreshold oscillations seen in the axons of cephalopods (modified from 1952).

To account for the precise values for a–h,[5] Hodgkin and Huxley devised the total current equation:

$$I = C_M dV/dt + G_K n^4 (V - V_K) + G_{Na} m^3 h (V - V_{Na}) + G_l (V - V_l)$$

In this equation, I is the total current crossing the membrane. That current has four components: the capacitative current $C_M dV/dt$, the potassium current $G_K n^4 (V - V_K)$, the sodium current $G_{Na} m^3 h (V - V_{Na})$, and the socalled "leakage current" $G_l (V - V_l)$, which is a sum of smaller currents for other ions. G_K, G_{Na} and G_l are the maximum conductance values for the different ionic currents. V is displacement of V_m from V_{rest}. V_K, V_{Na}, and V_l are the differences between equilibrium potentials for the various ions (that is, that voltage at which diffusion and the driving force of voltage are balanced so that there is no net current flow) and V_m. The capacitance, C_M, of the membrane can be understood as the ability of the membrane to store opposite charges on the intra- and extra-cellular sides. Finally, there are three coefficients, h, m, and n, whose values vary with voltage and time. The total current equation, as noted above, is derived from laws of electricity, such as Coulomb's law and Ohm's law (The latter is transparently represented in the component current equations for Na^+ and K^+).

Hodgkin and Huxley's primary accomplishment was to generate the equations for these variables and to determine the powers that they would take in the total current equation (see below). Hodgkin and Huxley show that each of a–h follows from the total current equation under specifiable conditions. For this contribution, they justly won the Nobel Prize.

Important as these equations are, Hodgkin and Huxley insist that the equations do not *explain* the action potential (see Bogen, 2005). They summarize decades of experiments. They embody a rich temporal constraint on any possible mechanism for the action potential. They allow neuroscientists to predict how current will change under various experimental interventions. They can be used to simulate the electrophysiological activities of nerve cells. They permit one to infer the values of unmeasured variables. And they constitute potent evidence that a mechanism involving ionic currents could possibly account for the shape of the action potential. However, according to Hodgkin & Huxley (1952), their model is not explanatory:

> The agreement [between the model and the voltage clamp data] must not be taken as evidence that our equations are anything more than an *empirical description of the time-course* of the changes in permeability to sodium and potassium. *An equally satisfactory description of the voltage clamp data could*

---

[5] This is how Hodgkin & Huxley (1952) write the equation. Contemporary textbooks use different formulations.

> *no doubt have been achieved with equations of very different form*, *which would* *probably have been equally successful in predicting the electrical behaviour of* *the membrane.* It was pointed out in Part II of this paper that certain features of our equations were capable of a physical interpretation, *but the success of the* *equations is no evidence in favour of the mechanism of permeability change that* *we tentatively had in mind when formulating them.* (1952, 541; italics added)

One could dismiss this curious passage as scientific modesty if it were not for the fact that Hodgkin and Huxley argue for their conclusions. Their arguments correspond closely to the distinctions raised in Sects. 2–4.

First, they insist that the equations provide nothing more than an empirical description of the time course of permeability changes (that is, the changes in conductance represented in the current equations for Na^+ and K^+). Here they are alluding to the variables n, m, and h. The current equation for K^+ involves the expression n^4. The current equation for Na^+ involves the expression m^3h. Discussing the equation for Na^+, Hodgkin and Huxley say that although they could have used a single variable, they found it easier to fit the curves with two. After the failure of one of their earlier hypothesized mechanisms, Hodgkin and Huxley realized that the techniques of electrophysiology did not suffice to pick out a uniquely correct characterization of the mechanism. Hodgkin reflects:

> We soon realized that the carrier model could not be made to fit certain results, for example the nearly linear instantaneous current voltage relationship, and that it had to be replaced by some kind of voltage-dependent gate. As soon as we began to think about molecular mechanisms it became clear that the electrical data would by themselves yield only very general information about the class of system likely to be involved. So we settled for the more pedestrian aim of finding a simple set of mathematical equations which might plausibly represent the movement of electrically charged gating particles. (Hodgkin, 1992)

Kenneth Cole, a collaborator of Hodgkin and Huxley, made the same point more dramatically when he said that the HH model merely "summarized in one neat tidy little package the many thousands of experiments done previous to 1952, and most subsequent ones" (1992, p. 151). The HH model (particularly its treatment of permeability changes) is in this respect more analogous to Ptolemy's planetary models, which neither involve nor imply any commitment to the existence of the epicycles, deferents, and equants from which they are constructed, than it is to Newton's gravitational model of planetary motion, which Newton presents to show how and why the planets move as they do. The equation embodies no commitments as to the mechanisms that change the membrane conductance, allow the ionic currents to flow, and coordinate them so that the action potential has its characteristic shape.[6] In the HH model, commitments about underlying mechanisms are replaced by mathematical constructs that save the phenomena a–h of the action potential much like Ptolemy's epicycles and deferents save the apparent motion of the planets through the night sky. The equations, in short, do not show *how* the membrane changes its permeability. As they said in 1952, the "Details of the mechanism will probably not be settled for some time" (1952, p. 504).

---

[6] I emphasize again that my focus, like Hodgkin and Huxley's, is on the treatment of permeability (conductance) changes, not on the change in V_m. Hodgkin and Huxley do show that changes in V_m can be explained in terms of permeability changes. (They have a proof to that effect.) But they have no explanation for the changes in permeability.

One might object to Hodgkin and Huxley's judgment on the grounds that the equations go beyond a mere description represents dependency relations among the items described by the equation. For example, the equations represent membrane conductance (permeability) as dependent upon voltage, and they describe sequential changes in currents across the membrane. This is true. But mathematical dependencies cannot be equated with causal or explanatory dependency relations. The equations must be supplemented by a causal interpretation: one might, for example, agree by convention that the effect variable is represented on the left, and the cause variables are represented on the right, or one might add "these are not mere mathematical relationships among variables but descriptions of causal relationships in which this variable is a cause and this other is an effect," and not vice versa, but the point is that one will have to specify which variables represent causes and which represent effects, and one will have to specify which of the myriad mathematical relationships contained in the equations are causal and which are mere correlations.[7] Mere correlations and effects of common causes can be represented as mathematical dependencies of one variable upon another, and equations can always be rewritten to put any variable that one likes on the left or the right. Absent an interpretation in terms of the underlying causal structure, such mathematical dependencies do not specify the causal dependencies that produce the time course of the action potential.

To be sure, Hodgkin and Huxley knew a good deal more about action potentials than is included explicitly in their mathematical model. Adding this detail helps to flesh out the mathematical model with details about a mechanism. Hodgkin and Huxley developed their model within a long tradition of electrophysiological research that had uncovered many of the components of the electrophysiological mechanisms in neurons. They knew that the action potential is produced by changes in membrane permeability. They knew that ions flux across the membrane toward their equilibrium potentials. They knew that this flux of ions constitutes a transmembrane current. This background sketch of a mechanism *does* provide a partial explanation for how neurons generate action potentials because it reveals some of the components of the mechanism, some of their properties, and some of their activities (it is a mechanism sketch). The HH equations supplement this background knowledge with explicit temporal constraints on the mechanism. The equations include variables that represent important components in the explanation. And they provide powerful evidence that a mechanism built from those components could possibly explain the action potential. And the equations, supplemented with a diagram of the electrical circuit in a membrane, and supplemented with details about how membranes and ion channels work, carry considerable explanatory weight. The equations without such interpretation—an interpretation that is difficult for those who know the mechanism of the action potential to imagine away—do not constitute an explanation. In order to explain, the equations of the model must be supplemented by an understanding of the mechanisms of the action potential—by an understanding of how the entities and activities in and around the membrane are organized together to produce the action potential.

---

[7] I am not making the absurd claim that no explanation can be represented in mathematical form. Equations are one convention among many for specifying causal relations. My point, rather, is that the mathematical expressions, as such, are consistent with a variety of different causal interpretations, many of which are spurious causal claims. The same equation allows one to represent the length of a pendulum as a cause or as an effect of its period, yet only the first gets the causal relationship right. The equation, absent causal interpretation, does not provide an explanation.

Hodgkin and Huxley insist that they have "no evidence" in favor of the mechanism that they "tentatively had in mind" when formulating their equations. According to that hypothesized mechanism, the membrane's permeability to Na^+ is regulated by the position of four particles in the membrane: three "activation molecules" that move from the outside of the membrane to sites on the inside, and one "inactivation molecule" that can block either the activation molecules or the flow of Na^+ through the membrane. The expression m^3h can then be interpreted as the joint probability that all three activation molecules are in their open state (with m being the probability that any one molecule has moved) and that no inactivation molecule is bound (h). When Hodgkin and Huxley say that they have no evidence for their hypothesized mechanism, they are referring to these variables in the current equations. The choice of a different strategy for building the equation (for example, using a single variable, or three rather than two) might suggest an entirely different physical interpretation (or mechanism) or none at all.

At most, this simple model of the activation and inactivation of sodium channels provides a "how-possibly" sketch of the action potential. Hodgkin and Huxley take an explicitly instrumentalist stance toward their model: "It was clear that the formulation we had used was not the only one that might have fitted the voltage clamp results adequately" (Huxley, 1963, p. 61). Indefinitely many equations could be used to predict the action potential's time-course. And these different mathematical equations might be given any number of biological interpretations such as the activation model sketched above. Hodgkin and Huxley had no reason to privilege this one how-possibly model above the others as a how-plausibly or how-actually model. To explain the action potential required further details about the molecular mechanisms underlying the permeability changes. Bertil Hille describes the origins of this research program:

> In the next decade, Clay and Armstrong and I began our independent research. In our first papers, we brought a clear list of 'molecular' assumptions to the table. They included the following ideas: ions are passing through aqueous pores that we called channels, ion channels are proteins, the channels for Na^+ and K^+ are different, they have swinging gates that open and close them, we can study their architecture by using electric currents to measure gating, permeation and block, and channel blockers are molecules that enter the pores and physically plug them. (Hille, Armstrong, & MacKinnon, 1999, p. 1106)

But at the time the idea of a channel was viewed with skepticism. It was merely a filler-term for an activity or mechanism to be named later:

> From 1965 to 1973, such ideas were debated annually at the meetings of the Biophysical Society. There, prominent scientists would routinely rise to request that anyone who chose to use the word "channel" avow first that it bears absolutely no mechanistic implications! It is probably fair to say that people thought the discussion about molecular mechanisms was premature. In 1969, when I had drafted a summary review of these ideas, Kenneth Cole, the dean of American biophysics, wrote to me: "I'm ... worried you may be pushing some of your channel arguments pretty far." (Hille et al., 1999, p. 1106)

The idea of activation molecules (let alone pores or gates) was at most a useful fiction—a how-possibly model—for Hodgkin and Huxley. It helped them to model the action potential, but it cannot be interpreted in terms of details about the membranes of nerve cells. Hille and his colleagues began to move beyond this useful

fiction by positing a set of how-possibly models and assessing them on experimental and theoretical grounds to produce a limited space of possible mechanisms. To leap to the end of the story, it is now well-known that conductance changes across the membrane are effected by conformation changes in ion-specific channels through the cell membrane. Biochemists have isolated these proteinaceous channels, they have sequenced their constituents, and they have learned a great deal about how they activate and inactivate during an action potential. It is in this wealth of detail (some of which is discussed below) about how these channels regulate the timing of the conductance changes, as described by the HH equations, that explain the temporal course of the action potential.

Only with the discovery of these molecular mechanisms could the action potential be not merely modeled but explained. As Michael Mauk notes:

> There was little to be learned from the particular mathematical implementation that H[odgkin] and H[uxley] used to represent voltage-dependent conductances. Because they were intended only as mathematical tools to produce the correct input/output behavior, the ingredients of the [phenomenological] model did not need to reflect the underlying biological processes. For example, the conductances could have been described in lookup tables. Thus, like experiments with only one possible outcome, the ability to build these [phenomenological] models meant little mechanistically. (2000, p. 650)

The Hodgkin and Huxley model illustrates the normative distinctions of Sects. 2–4. At least certain aspects of the model (specifically, the equations governing the values of n, m, and h) are merely phenomenological models.[8] These equations characterize how specific ion conductances change with voltage and time, but they do not explain why they change as they do, when they do. Without an account of the underlying mechanisms of the conductance change, the buck of accounting for the temporal features of the action potential (as specified in a–h) is merely passed on to some-conductance-changing-process-we-know-not-what. This filler-process was later completed as Hille and his colleagues investigated the structure and function of ion channels. This allowed them to weed merely how-possibly mechanisms out of the space of plausible mechanisms.

6 Evaluating mechanistic explanations

Models are explanatory when they describe mechanisms. Perhaps not all explanations are mechanistic. In many cases, however, the distinction between explanatory and non-explanatory models is that the latter, and not the former, describe mechanisms. It is for this reason that such models are useful tools for controlling and manipulating phenomena in the world.

Mechanistic models are *ideally complete* when they include all of the relevant features of the mechanism, its component entities and activities, their properties, and their organization. They are *pragmatically complete* when they satisfy the pragmatic demands implicit in the context of the request for explanation. What follows is an admittedly preliminary and incomplete checklist for assessing mechanistic

---

[8] It turns out, in retrospect, that aspects of the equations for these conductance changes do correspond to features of the ion channels, but this is, as Hodgkin and Huxley would have noted, not a perfect fit and, at any rate, is merely fortuitous.

explanations, for distinguishing how-possibly models from how-actually explanations (see Craver, forthcoming; Glennan, 2005 for further details), and for shrinking the space of plausible mechanisms (see also Craver & Darden, 2001).

6.1 The phenomenon

As Stuart Kauffman (1971) and Stuart Glennan (1996, 2002) argue, mechanisms are always mechanisms *of* a given phenomenon. They are the mechanisms *of* the things that they *do*. The mechanism of the action potential generates action potentials. The core normative requirement on mechanistic explanations is that they must account fully for the explanandum phenomenon. As such, a mechanistic explanation must begin with an accurate and complete characterization of the phenomenon to be explained.

Phenomena are typically *multifaceted*. Part of characterizing the action potential phenomenon involves noting that action potentials are produced under a given range of *precipitating conditions* (for example, a range of depolarizations in the cell body or axon hillock). But, as Hodgkin and Huxley's a–h illustrate, there is much more to be said about the *manifestations* of an action potential. It is necessary to describe its rate of rise, its peak magnitude, its rate of decline, its refractory period, and so on. The action potential is characterized by a number of input–output relationships, each of which must be satisfied by any explanatory model of the mechanism. Furthermore, neuroscientists' understanding of the action potential has expanded considerably since 1952. A how-possibly model that accounts for features a–h, but not the subsequent discoveries concerning action potentials, would be a mere how-possibly model. It would not explain the action potential.

Second, it is insufficient to characterize the phenomenon only under standard precipitating conditions. A complete characterization of the phenomenon requires one to know its *inhibiting conditions*—that is, the conditions under which the phenomenon fails to occur. Action potentials can be prevented, for example, by applying tetrodotoxin (TTX), which blocks the flow of Na^+ through Na^+ channels, or by removing Na^+ from the extracellular fluid. If one truly understands the mechanisms of the action potential, one should be able to say why they are *not* produced under these conditions.

Third, a complete characterization of the phenomenon requires knowing the phenomenon's *modulating conditions*—that is, knowing how variations in background conditions alter the action potential. For example, one wants to know how the action potential will change if one were to change the neuron's diameter, or the density of ion channels in a given stretch of membrane, or the extracellular concentration of Na^+.

Fourth, one has not fully characterized the action potential unless one also knows how it behaves under a variety of *non-standard conditions*. Most laboratory conditions are non-standard. If one connects a squid giant axon (the experimental system in which Hodgkin and Huxley's experiments were performed) to a space clamp or a voltage clamp (crucial experimental innovations in this historical episode), one observes the behavior of cells under conditions that would never occur in a normal organism. Although such experiments are not physiologically relevant (that is, relevant to the behavior of neurons in a normal cell under standard operating conditions), they are nonetheless part of how the mechanism works if manipulated in specific ways. Two how-possibly mechanisms can account equally well for the

capacity of a neuron to produce standard action potentials under physiologically normal precipitating conditions but nonetheless diverge considerably in their ability to account for features of action potentials in inhibiting, modulating, and otherwise non-standard conditions.

Fifth, a variety of *byproducts* or side-effects of the phenomenon can also be crucial for sorting how-possibly from how-actually models and sketches from complete mechanistic models. Byproducts include a range of possible features that are of no functional significance for the phenomenon (for example, they do not play any role in a higher-level mechanism) but are nonetheless crucial for distinguishing mechanisms that otherwise account equally well for the phenomenon (see Cummins, 2000, pp. 123–124). The activation of Na^+ channels, for example, is accompanied by a gating charge, a very slight movement of charges across the membrane. Why is there a gating charge? According to the standard textbook model, the activation of Na^+ channels involves rotating an α-helix, which is composed of regularly spaced positive charges. In fact, it turns out that the gating current is precisely equal to the amount of charge moved across the membrane as the α-helix rotates. All of the current competing models of voltage sensor are designed to accommodate the gating charge (see Swartz, 2004). Gating charge apparently plays no role in the electrical activities of nerve cells, but it is nonetheless an aspect of the voltage sensor, and it is one that any how-actually model has to account for.

In summary, mechanistic explanations can fail because one has mischaracterized the phenomenon, or because one has only partially characterized the phenomenon to be explained. One can conjecture a mechanism that adequately accounts for some narrow range of features of the phenomenon but that cannot accommodate the rest. For this reason, descriptions of multiple features of a phenomenon, of its precipitating, inhibiting, modulating, and non-standard conditions, and of its byproducts, all constrain mechanistic explanations and help to distinguish how-possibly from how-actually explanations. Similarly, mechanism sketches, with large gaps and question marks, might explain some aspects of the explanandum phenomenon but fail to explain others. Hodgkin and Huxley's background sketch explains the shape of the action potential in terms of changes in currents, but the sketch does not explain the conductance changes that lie at the heart of the model. To characterize the phenomenon correctly and completely is the first restrictive step in turning a model into an acceptable mechanistic explanation.[9]

6.2 Parts

Mechanistic explanations are *constitutive* explanations: they explain the behavior of the mechanism as a whole in terms of the organized activities of and interactions among its *components*. Components are the entities in a mechanism—what are commonly called "parts." Action potentials are explained by appeal to components such as Na^+ and K^+ channels, ions, and protein chains.

---

[9] One way to confirm that one has properly characterized a phenomenon is to see whether a system that embodies the phenomenon can be inserted back into a higher-level system without disturbing the behaviors of that higher-level system. A compelling example of this sort of work involves the use of "hybrid models," which causally insert simulations of a part's behavior into a biological system to see if the system properties are preserved if the part behaves as the simulation demands (see Prinz, 2004). For example, one might insert an artificial neuron, that is, a simulation, into an actual neural system to see if the simulation does what it needs to do for the neural system to do what it does the way that it normally does it.

What I stress, however, is the difference between models that describe the parts of a mechanism and those that posit relationships among useful fictions that fail to correspond to parts of the implementing mechanism. No neuroscientist would claim, for example, that it makes no difference to the explanation of the action potential whether ions move across the membrane by active transport, passive diffusion, or a mechanism made of Swiss cheese (to pick a philosophically charged example). One might be entertained by building a model of the action potential out of Swiss cheese, and it would be impressive indeed if this model could reproduce the form of the action potential, but no reputable journal would publish the model, let alone allow the author to claim that it counted as an explanation of the action potential. Neurons are not made of Swiss cheese. Nor are they made of Hodgkin and Huxley's "activation particles" that move within the membrane and change its conductance. Activation particles are fictional constructs, and although functional relationships among the activation particles can account for all of the features of an action potential, they do not explain the action potential. Similarly, box-and-arrow diagrams can depict a program that transforms relevant inputs into relevant outputs, but if the boxes and arrows do not correspond to real component entities and activities, one is providing a redescription of the phenomenon (such as the HH model) or a how-possibly model (such as Hodgkin and Huxley's working model of conductance changes), not a mechanistic explanation.

To distinguish good mechanistic explanations from bad, one must distinguish real components from fictional posits. There is no clear evidential threshold for saying when one is describing real components as opposed to fictional posits. Nonetheless, the following criteria are satisfied by real parts and can be used to distinguish how-possibly from how-actually explanations.

First, we say that parts are real when they exhibit a *stable cluster of properties* (see Boyd, 1999). Hille's speculative channels were gradually transformed into stock-in-trade entities as it became possible to sequence them, recover their secondary and tertiary structure, describe their interactions with chemical agonists and antagonists, characterize their voltage-dependence and rapid inactivation, and so on. As details mounted about the shapes of the channels, their components, their causal powers, and their subtypes, it became increasingly difficult to dismiss channels as merely a hypothesis being "pushed too far."

Second, and related, we are confident the parts are real when they are *robust* (Wimsatt, 1981), that is, detectable with multiple causally and theoretically independent devices.[10] The convergence of multiple lines of independent evidence about Na^+ channels convinced neuroscientists of their existence. Ion channels can be isolated from the membrane, purified, and sequenced. Their behavior can be detected en masse through intra- and extracellular recording techniques, and they can be monitored individually with single-channel patch-clamp techniques. They can be manipulated with pharmacology, they can be altered with site-specific mutagenesis, they can be crystallized and X-rayed, and they can be seen through an electron microscope. Using multiple techniques and theoretical assumptions to reason to the existence of a given item decreases the probability that conclusions drawn from any single technique or mode of access are biased or otherwise faulty (Psillos, 1999, Salmon, 1984).

---

[10] This sentiment is captured by the adage that all techniques stink, but they stink in different ways.

Third, we know that parts are real when we can use them to *intervene* into other components and activities (Hacking, 1983). It should be possible, that is, to manipulate the entity in such a way as to change other entities, properties, or activities. One can manipulate Na^+ channels to alter the membrane potential, to change Na^+ conductance, to open K^+ channels, or to balance current.

Fourth, the components should be plausible-in-the-circumstances or, for physiological mechanisms, *physiologically plausible*. They should not exist only under highly contrived laboratory conditions or in pathological states unless one is interested only in explaining the behavior of the mechanism in those highly contrived or pathological states. What constitutes a contrived condition or a pathological state varies across explanatory contexts. If one is trying to explain healthy functions, then pathological conditions might be considered physiologically implausible. If, on the other hand, one is trying to explain a disease process, one's explanation might be physiologically implausible if it assumes conditions only present in healthy organisms. What matters is that the parts' existence should be demonstrable under the conditions relevant to the given request for explanation.

Finally, the components must be *relevant* to the phenomenon to be explained. Some parts of the cell are relevant to the action potential and some are not. Sodium and potassium channels are clearly relevant to the action potential; they are part of the mechanistic explanation for how action potentials are generated, why they have their characteristic shapes, and so on. Vesicles in the axon terminal, the nuclear membrane, and DNA are not relevant to these features of the action potential. One way to establish that a component is relevant is to intervene to alter or delete the component (e.g., to pharmacologically inactivate Na^+ channels) and to observe changes in the behavior of the mechanism as a whole (e.g., the rising phase of the action potential).[11] Another way is to intervene to change the behavior of the mechanism as a whole (e.g., by injecting current into a cell) and to observe the behavior of the component parts (e.g., the conductance changes in ion channels). These experiments, when they are conducted properly and with proper controls, establish a symmetrical counterfactual dependence relationship between components, properties, and activities and the behavior of the mechanism as a whole (for more details, see Craver, forthcoming, Chap. 4).

This is neither an exhaustive list of criteria nor an exhaustive discussion of the items in it. Nonetheless, in making these criteria explicit, I take steps toward spelling out when one is justified in presuming that one has moved beyond providing merely a how-possibly account or a filler term, and toward describing an actual mechanism. Hille and Armstrong's channel hypotheses moved from a how-possibly posit to a how-actually description of a mechanism as findings about membrane-spanning ion channels satisfied the above criteria.

6.3 Activities

Mechanisms are composed of entities and *activities*. Activities are the things that entities do. For example, Na^+ channels activate and inactivate, and ions diffuse down their concentration gradients. Activities are the causal components of mechanisms.

It will not do to describe the activities in mechanisms as merely input–output pairs, for there can be input–output pairs that are not explanatory. One can use

---

[11] All of these experiments are subject to well-known problem-cases, but there are also well-known ways of dealing with those problem cases, in part by using multiple techniques at once.

input–output pairs to describe non-causal temporal sequences (input crowing roosters, output dawn), effect-to-cause pairs (input refractory period of the action potential, output rising phase), correlations between the effects of a common cause (input falling barometer, output storm), and irrelevant pseudocause-to-effect pairings (input blessing, output action potential) (see Craver, forthcoming, Chap. 3). It will not help matters to require that the input–output regularity support counterfactuals (as Weber, 2005 requires in his discussion of the mechanisms of the action potential), because not all counter-factual supporting generalizations are explanatory. If the rooster were to be crowing, dawn would be coming. If my barometer were falling, a storm would be on the horizon. (See Lewis's 1973 distinction between backtracking and non-backtracking counterfactuals.) Clearly, the requisite notion of an activity must be more restrictive than an input–output pair that sustains counterfactuals.

One crucial requirement on activities can be expressed as a restriction on the kind of input–output relationships that can count as explanatory. Following Pearl (2000) and Woodward (2003), I propose that the activities in mechanisms should be understood partly in terms of the ability to manipulate the value of one variable in the description of a mechanism by manipulating another. If there is an activity connecting some feature of a mechanism X to another feature of the mechanism Y, then it should be possible to manipulate Y by manipulating X. To say that depolarizing the membrane activates Na^+ channels is, in part, to say that one can change the conductance of the Na^+ channels by manipulating the voltage across the membrane. To say that a ball on the end of a protein chain inactivates Na^+ channel is to say that one could manipulate the conductance of Na^+ channels by manipulating the ball and chain (for example, that one could prevent Na^+ inactivation by cutting the chain or changing the size and shape of the ball).

There are a number of theoretical virtues that come from thinking of activities in this way (see Pearl, 2000, Woodward, 2003). For present purposes, I merely want to point out that one can use this manipulability criterion as a test for distinguishing causally relevant from causally irrelevant factors. One cannot make the sun rise by intervening into a rooster's suprachiasmatic nucleus. One cannot change the action potential by intervening on its refractory period. In Balmer's formula, there is no independent variable to manipulate: it is merely an empirical hypothesis. The requirement of manipulability thus has the ability to distinguish backtracking from non-backtracking counterfactuals and to distinguish explanatory from non-explanatory input–output pairs.

One will perhaps have noticed that Snell's law does satisfy this requirement of manipulablity. If one wants to know why a laser bends *by a certain* angle as it passes from water to glass, Snell's law shows how this variable depends upon the angle of incidence and the refractive indices involved. However, if one wants to understand why light bends as it passes from one medium to another, then Snell's law is merely a phenomenal description: it describes nothing of the constitutive mechanisms by which light exhibits these behaviors. Likewise, the Hodgkin and Huxley equations can be used to show how certain variables depend systematically on others, but does not say why those relationships hold. For that, one needs an understanding of the underlying mechanisms. This will detail the lower-level activities responsible for the non-backtracking relationships expressed in the Hodgkin and Huxley model.

6.4 Organization

The last crucial feature of mechanistic explanations is the organization of the components such that they jointly exhibit the phenomenon to be explained. Mechanistic explanations are not merely aggregative; the phenomenon is not merely a sum of properties of the component parts. One cannot interchange the parts of a mechanism indiscriminately without disturbing the behavior of the whole. One cannot typically add or remove parts without producing discontinuities in the behavior of the whole. These tests reveal symptoms indicating that the behavior of the mechanism as a whole depends upon how the components and their activities are arranged spatially, temporally, and hierarchically (see Wimsatt, 1997).

Contrast this view with Cummins's (1975, 1983, 2000) idea that organization must be something that can be specified in a flow chart or drawn in a box-and-arrow diagram. Mechanistic explanations are frequently presented in diagrams that show how one stage of the mechanism is productively continuous with its predecessor. However, Cummins does not comment on the fact that such diagrams, while widely recognized as useful heuristics, often provide only the illusion of understanding a mechanism. Boxes labeled with filler terms take on the appearance of real components in the mechanism. Arrows between boxes can be used to hide "some-process-we-known-not-what" that plays a crucial role in a mechanism. And finally, anything can be represented as a box-and-arrow diagram. It is a flexible representational format and so places no meaningful constraints on mechanistic explanations.

In the explanation for the action potential, different forms of spatial and temporal organization are primarily important. It matters, for example, that the axon hillock connecting the cell body to the axon is very dense in Na^+ channels. It matters that the ion channel has a configuration that opens into a pore, and that the protein ball is large enough to obstruct the channel. It matters how different charges are distributed through the ion channel and how they are oriented with respect to one another and to the electrical fields produced across the membrane. It matters when the different channels activate, how long they stay open, and when they inactivate. To provide a mechanistic explanation, one shows how the different features of the phenomenon depend upon the organizational features of the underlying mechanism.

A great deal more can be said about the organization of mechanisms (see Craver, forthcoming), but these brief remarks should suffice to show that it is insufficient merely to describe components and their activities. One has to further describe how those entities and activities are organized (e.g., spatially and temporally) into a mechanism. How many forms of organization there are, and which forms of organization can legitimately appear in a mechanistic explanation, are questions left for another discussion.

7 Conclusion

The historical example of the Hodgkin and Huxley model of the action potential, and the subsequent development of an explanation for the action potential, illustrates that models are often not explanatory. Explanations are supposed to do more than merely predict how the target mechanism will behave under a variety of conditions. They are, in addition, supposed to account for all aspects of the phenomenon to be explained

by describing how the component entities and activities are organized together such that the phenomenon occurs. Mechanistic models explain.

One goal for building a philosophical analysis of scientific explanation, and for thinking carefully about the role of models in science, is to clarify the normative constraints at work in the construction, evaluation, and revision of scientific explanations. I have shown that in one area of neuroscience, mere models are not accepted as explanations. I have argued that this assessment is correct by appealing to the fact that mere models are of little help in thinking about how to control the behavior of a mechanism. The ability to control a system is a mark of understanding how that system works: one with that ability will be able to answer a range of questions about how the system would behave if it were to be altered in any number of ways or placed in any number of conditions (see Woodward, 2003). If this is a reasonable view of what constitutes a complete explanation, and if this view helps to reveal the difference between adequate and inadequate explanations, then one should reflect on whether the autonomous explanations championed in other parts of cognitive neuroscience really can afford to remain so autonomous. Phenomenal models, while sometimes appropriate in a given conversational context, are at best shallow explanations.

I have perhaps said enough to raise the hackles of those who promote decompositional, box-and-arrow, explanations in many areas of cognitive science that describe functions abstracted away from the details of their realizing mechanisms. I have argued that in many areas of science, constitutive explanations of this sort are treated as unsatisfactory precursors to explanations, and I have argued that such functional explanations frequently allow one to answer fewer what-if-things-had-been different questions than one that, in addition, includes details about the state of the underlying mechanism. This by no means entails that there are no higher-level explanations or that higher-level phenomena are never relevant to a given phenomenon, though it is beyond the scope of this paper to argue for this thesis. It does entail, however, that good constitutive explanations go beyond data summaries, how-possibly models, and sketches to provide a detailed description of the relevant components and activities constituting an actual mechanism.

References

Boyd, R. (1999). Kinds, complexity, and multiple Realization. *Philosophical Studies, 95*, 67–98.
Bogen, J. (2005). Regularities and causality; generalizations and causal explanations. *Studies in the History and Philosophy of Science*, C *36*, 397–420.
Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT Press.
Cole, K. (1992). Neuromembranes: Paths of ions. In I. F. G. Worden, J. P. Swazey, & G. Adelman (Eds.), *Neurosciences, paths of discovery*, I. Cambridge, Mass.: MIT Press.
Craver, C. F. (forthcoming). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience.* Oxford: Oxford University Press.
Craver, C. F., & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In P. K. Machamer, R. Grush, & P. McLaughlin (Eds.), *Theory and method in the neurosciences.* (pp. 112–137). Pittsburgh, PA: University of Pittsburgh Press.
Cummins, R. (1975). Functional analysis. *Journal of Philosophy, 72*, 741–765.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, MA: Bradford/ MIT Press.

Cummins, R. (2000). "How does it work?" vs. "What are the laws?" Two conceptions of psychological explanation. In F. Keil, & R. Wilson (Eds.), *Explanation and cognition* (pp. 117–145). Cambridge, MA: MIT Press.

Dennett, D. C. (1994). Cognitive science as reverse engineering: Several meanings of 'Top Down' and 'Bottom Up'. In D. Prawitz, B. Skyrms, & D. Westerståhl (Eds.), *Logic, methodology and philosophy of science IX* (pp. 679–689). Amsterdam, North-Holland: Elsevier Science, BV.

Feree, T., & Lockery, S. R. (1999). Computational rules for chemotaxis in the nematode *C. elegans*. *Journal of Computational Neuroscience, 6*, 263–277.

Giere, R. (1999). *Science without laws*. Chicago, IL: University of Chicago Press.

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

Glennan, S. S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69* (Supplement), S342–S353.

Glennan, S. S. (2005). Modeling mechanisms. *Studies in the History and Philosophy of Science, C. 36*, 443–464.

Hacking, I. (1983). *Representing and intervening*. Cambridge, U.K.: Cambridge University Press.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Hille, B., Armstrong, C., & MacKinnon, R. (1999). Ion channels: From idea to reality. *Nature Medicine, 5*, 1105–1109.

Hodgkin, A. L. (1992). *Chance & design: Reminiscences of science in peace and war*. Cambridge: Cambridge University Press.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology, 117*, 500–544.

Huxley, A. F. (1963). The quantitative analysis of excitation and conduction in nerve. At Nobelprize.org. http://nobelprize.org/medicine/laureates/1963/huxley-lecture.html

Kauffman, S. A. (1971). Articulation of parts explanation in biology and the rational search for them. In R. C. Buck, & R. S. Cohen (Eds.), *PSA 1970*. Dordrecht: Reidel.

Kitcher, P. (1984). 1953 and all that: A tale of two sciences. *Philosophical Review, 93*, 335–373.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*, 556–567.

Lycan, W. (1999). The continuity of levels of nature. In W. Lycan (Ed.), *Mind and cognition: A reader*, 2nd edn. Malden, MA: Blackwell Publishers.

Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 57*, 1–25.

McClelland, J., & Rumelhart, D. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 2. Cambridge, MA: MIT Press.

Morgan, M. S., & Morrison, M. (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge: Cambridge University Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Piccinini, G. (forthcoming). Computational modeling vs. computational explanation: is everything a turing machine, and does it matter to the philosophy of mind? *Australasian Journal of Philosophy*.

Povinelli, D. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works*. Oxford: Oxford University Press.

Prinz, A. (2004). Neural networks: Models and neurons show hybrid vigor in real time. *Curr Biol, 16*, R661–R662.

Psillos, S. (1999). *Scientific realism: How science tracks truth*. Routledge.

Rummelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge: MIT Press.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Simon, H. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Suppe, F. (1989). *The semantic conception of theories and scientific realism*. Urbana, IL: University of Illinois Press.

Suppes, P. (1967). What is a scientific theory? In S. Morgenbesser (Ed.), *Philosophy of science today*. New York: Basic Books.

Swartz, K. J. (2004). Towards a structural view of gating in potassium channels. *Nature Reviews Neuroscience, 5*(12), 905–916.

Weber, M. (2005). *Philosophy of experimental biology*. Cambridge: Cambridge University Press.

Wimsatt, W. (1981). Robustness, reliabilty, and overdetermination. In M. Brewer, & B. Collins (Eds.), *Scientific inquiry and the social sciences*. San Francisco, CA: Jossey-Bass Publishers.

Wimsatt, W. (1997). Aggregativity: Reductive heuristics for finding emergence. In L. Darden (Ed.), *PSA-1996*, v.2. *Philosophy of Science, 66*, S372–S384.

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