or what the available metaphysical options are. Even if the manipulationist view does not identify the truth-maker for causal claims, it is nonetheless an illuminating analysis of the causal truths themselves, and it is crucial for the project of deciding which putative metaphysical explanations (that is, which truth-makers) are adequate and which are not.

Although I display some of the merits of the manipulationist approach relative to some competitors (mechanical and transmission accounts), I do not argue that one can make sense of causal relevance only by appeal to manipulability relations. I do not rule out the possibility that (E1)–(E5) might be satisfied by other accounts of causation. Nor do I rule out the possibility that there is more to learn about causation by investigating such alternatives. I believe, for example, that Hitchcock's comparative conception of the statistical dependency relations involved in causation (Hitchcock 1996) can help to remove certain ambiguities in the manipulationist approach (I build on this idea in Chapter 6). I believe further that the notion of "productive activities" developed by Machamer et al. (2000) and deployed by Craver and Darden (2001) and Darden and Craver (2003) is extremely useful for describing the history of science, for understanding aspects of scientific change, for thinking about how to build explanations, and for thinking about the metaphysics of causation (for a discussion of this issue, see Tabery 2004). Nonetheless, I now have a view of causal and explanatory relevance that can resolve some of the problems that plague the CL model, the U-model, and the PDP model. This seems to me a very friendly amendment to many current mechanistic views of etiological explanation, including my own (Machamer et al. 2000; Craver 2001). By supplementing the account of mechanisms in this way, one adds a normative dimension, showing what it means to correctly identify causally relevant factors within a mechanism. In the next chapter, I show how this view of causal relevance can be embedded within an account of mechanisms and can be extended to provide an account of *constitutive* explanatory relevance.

4

The Norms of Mechanistic Explanation

Summary

In this chapter, I develop a causal-mechanical model of constitutive explanation. The account satisfies two goals: first, to provide an alternative to classical reduction for thinking about constitutive explanation, and second, to show how the systems tradition (exemplified by Cummins's view of explanation as functional analysis) would have to be amended and revised if it is to offer a normatively adequate account of constitutive mechanistic explanation. I build my account by considering the discovery of the mechanism of the action potential and the diverse kinds of experiment required to show that a component is relevant to such a mechanism. The resulting view is a causal-mechanical competitor to reduction as a way of understanding interlevel relationships in neuroscience and beyond.

1. Introduction

Explanations in neuroscience describe mechanisms. Some mechanistic explanations are etiological; they explain an event by describing its antecedent causes. Dehydration is part of the etiological explanation of thirst. Prion proteins are part of the etiological explanation of Creutzfeldt-Jacob disease. Excessive repetition of the CAG nucleotide pattern on the fourth chromosome is part of the etiological explanation for Huntington's

disease. Other mechanistic explanations are constitutive or componential;[1] they explain a phenomenon by describing its underlying mechanism. The NMDA receptor is part of the constitutive explanation of LTP. The hippocampus is part of the constitutive explanation for spatial memory. Ions are part of the constitutive explanation for the action potential. In this chapter, I develop a normatively adequate account of constitutive explanation.

There are two dominant and broad traditions of thought about constitutive explanation: the reductive tradition and the systems tradition. My view is a development and elaboration of one strand in the systems tradition.

The reductive tradition construes constitutive explanation as a species of CL explanation holding between theories at different levels. The explanation proceeds by constructing identity statements (or partial identity statements) between the kind-terms of the higher-level theory and those of the lower-level theory and then deriving the laws of the higher-level theory from the laws of the lower-level theory. The derivational requirement serves two purposes. First, it provides an epistemic account of explanation, according to which understanding is *rational expectation* of the *explanandum* on the basis of the *explanans* (in accordance with the nomic expectability thesis discussed in Chapter 2). Second, it offers a *regulative ideal* for explanation. If the explanation is ideally complete, one should literally be able to derive the *explanandum* from the *explanans*. Even if few explanations in neuroscience or elsewhere live up to this standard, the reduction model nonetheless provides a clear statement of what is required of an adequate explanation.

Although most philosophers of neuroscience (including John Bickle 1998, 2003; P. S. Churchland 1986; and Schaffner 1993a and b) fall in the reductive tradition, this classic view of reduction has few remaining advocates among philosophers of mind and philosophers of science. The most cited reason is that it is impossible to formulate the requisite identities because higher-level kinds are multiply realized by lower-level kinds to such an extent that there is no question of forming identities between the kind-terms in the higher-level theory and those in the lower-level theory. The conceptual taxonomies at different levels are askew, and therefore the one-to-one mapping that reduction requires is unlikely to fit the facts.

[1] I borrow the term "constitutive" from Salmon (1984). I mean by "constitutive" a relationship between the behavior of a mechanism as a whole and the organized activities of its individual components. I understand that the word "constitutive" is used for other purposes in metaphysics, but I am following Salmon's usage.

A second reason why reduction is unpopular is that real explanations in neuroscience look nothing like the explanations that the reduction model requires. Defenders of reduction have been forced to endorse the limited claim that the model serves mainly as a "regulative ideal" that is entirely "peripheral" to the practice of biology and neuroscience (see Schaffner 1974; P. M. Churchland 1989). A third, and least cited, reason why reduction has few supporters is that the deductive model of explanation on which reduction is premised has the varied shortcomings I discuss in Chapter 2. It is not sufficient to explain a theory merely to be able to derive it from another theory. The required derivation can be constructed on the basis of mere correlations, temporal sequences, effect-to-cause generalizations, and incomplete explanations.

The systems tradition, in contrast, construes explanation as a matter of decomposing systems into their parts and showing how those parts are organized together in such a way as to exhibit the *explanandum phenomenon*. In this tradition, I include philosophers of biology and psychology who discuss explanation by functional analysis (Fodor 1968; Cummins 1975, 1983, 2000), by decomposition (Simon 1969; Wimsatt 1974; Haugeland 1998), by identifying homunculi (Fodor 1968; Dennett 1978; Lycan 1987), by reverse engineering (Dennett 1994), by taking the design stance (Dennett 1987), by describing the articulation of parts (Kauffman 1971), and, finally, by discovering mechanisms (Bechtel and Richardson 1993; Glennan 2002). Lycan, following Dennett (1978), describes this form of explanation with the metaphor of little men:

> We explain the successful activity of one homunculus not by idly positing a second homunculus within it that successfully performs the activity, but by positing a team consisting of several smaller, individually less talented and more specialized homunculi—and detailing the ways in which the team members cooperate in order to produce their joint or corporate output. (Lycan 1987; in Lycan 1999: 51)

Dretske mobilizes several different metaphors in his article, "If you can't make one, you don't know how it works":

> All I mean to be suggesting by my provocative title is something about the spirit of philosophical naturalism. It is motivated by a constructivist's model of understanding. It embodies something like an engineer's ideal, a designer's vision, of what it takes to really know how something works. You need a blueprint, a recipe, an instruction manual, a program. (Dretske 1994: 468)

Cummins's account of explanation by functional analysis is the most rigorous formulation of the systems tradition prior to recent discussions of mechanisms and mechanistic explanation. I reference Cummins's account throughout this chapter. On his view, the *explanandum* is some capacity ψ of a system S. S's ψ-ing is explained by analyzing it into subcapacities $\{\phi_1, \phi_2, \ldots, \phi_n\}$ and showing that ψ is produced through the programmed exercise of the subcapacities. To show that ψ can be produced, in this sense, through the programmed exercise of the subcapacities, one specifies a box-and-arrow diagram showing how the subcomponents work together such that they ψ. For example, the capacity of the neuron (S) to generate action potentials (ψ) would presumably be explained by a box-and-arrow diagram that exhibits the programmed exercise of such capacities as rotating, changing conformation, and diffusing.

As Cummins's account illustrates, systems explanations involve showing how something works rather than showing that its behavior can be derived from more fundamental laws (Dretske 1994; Cummins 2000; Bechtel and Abrahamsen 2005). This view of explanation has several advantages over the model required by the reductive tradition. For example, it does not matter for the systems tradition that the *explanandum phenomena* might be multiply realized. If the multiple realizability of nonfundamental phenomena raises a problem for classical reduction (there is debate on this matter), it is because classical reduction requires the translation of kind-terms in one theory into those of another theory. Translation is not required because the systems tradition rejects the idea that explanations are arguments. All that matters is that the phenomenon is realized by some underlying mechanism. Furthermore, systems explanations are not peripheral to the practice of neuroscience; they are much more accurate descriptions of neuroscientific explanations than the reduction model supplies. Finally, systems explanations need not inherit the limitations of the CL model; they promise an altogether different vision of scientific explanation.

But what, exactly, is that alternative vision of explanation? What does it mean to "know how something works" or to "reduce a capacity to the programmed exercise of sub-capacities" if not that one can derive the behavior of a mechanism as a whole from the organized behaviors of its parts? What distinguishes good constitutive explanations from bad? What does it mean to have a complete systems explanation? How does one decide which parts should be included in a systems explanation and which parts are irrelevant? If the systems tradition is to present a complete alternative to reduction, then it must provide an alternative set of norms by which explanations should be assessed. Otherwise, it provides an adequate surface description of constitutive explanations in neuroscience without challenging the core idea of explanation underlying classical reduction.

In this chapter, I construct a normatively adequate *mechanistic* model of constitutive explanation (henceforth, mechanistic explanation). Chapter 1, Section 2 contains a sketch and overview of my basic position. Those wanting merely a summary should consult that sketch and the conclusion of this chapter. The primary purpose of this chapter is to move beyond that sketch and to show how the simple idea of explanation by decomposition can be made precise and normatively rigorous. In doing so, I present a more detailed and elaborate exposition of the systems tradition than is currently available, and I provide it with the tools to challenge reduction as a normative model of constitutive explanation in neuroscience and beyond.

I construct my model of mechanistic explanation to serve two ends: (1) to distinguish how-possibly explanations from how-actually explanations, and (2) to distinguish mechanism sketches from mechanism schemata. These distinctions are introduced in Section 2. In Section 3, I illustrate progress along these two dimensions by considering how neuroscientists moved beyond the Hodgkin and Huxley model of the action potential in order to provide a complete explanation of the conductance changes constituting the action potential. In subsequent sections, I show how Cummins's account of functional analysis (and so the systems tradition generally) can be supplemented and transformed to become an account of mechanistic explanation that rivals reduction as a regulative ideal for explanation. The regulative ideal is that constitutive explanations must describe all and only the component entities, activities, properties, and organizational features that are relevant to the multifaceted phenomenon to be explained. I build my account slowly by considering separate aspects of mechanistic explanation sequentially. These include:

(i) the nature of the *explanandum phenomenon* (Section 4);
(ii) the constitutive relationship between a phenomenon and its components (Section 5);

(iii) the difference between real components and useful fictions (Section 5);
(iv) the nature of capacities or activities (Section 6);
(v) the nature of mechanistic organization (Section 7); and
(vi) the nature of constitutive explanatory relevance (Section 8).

This last topic—the problem of saying what it means for a component to be explanatorily relevant to a phenomenon—has thus far been entirely neglected by both the systems tradition and the reduction tradition. In Section 8, which could be considered a chapter within this chapter, I introduce this problem and offer a causal-mechanical solution.

2. Two Normative Distinctions

Throughout this chapter, I am guided by two normative distinctions that are implicit in the practices of constructing and evaluating mechanistic explanations. An adequate account of mechanistic explanation should help one to understand how these distinctions are drawn.

First, the proposed account should have resources adequate to distinguish how-possibly models from how-actually models. *How-possibly models* have explanatory purport, but they are only loosely constrained conjectures about the sort of mechanism that might suffice to produce the *explanandum phenomenon*. They describe how a set of parts and activities might be organized together such that they exhibit the *explanandum phenomenon*. One might have no idea if the conjectured parts exist or, if they do, whether they are capable of engaging in the activities attributed to them in the model. Some computer models are purely how-possibly models. For example, one might simulate aspects of the visual system in LISP without any commitment to the idea that the brain is somehow executing CARs and CDRs (the basic operations of LISP) through its neural networks. How-possibly models are often heuristically useful in constructing and exploring the space of possible mechanisms, but they are not adequate explanations. *How-actually models*, in contrast, 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 these extremes is a range of *how-plausibly* models that are more or less consistent with the known constraints on the components,

their activities, and their organization.[2] To continue Dennett and Lycan's metaphor for functional analysis, without some restrictions on who can be a homunculus, on which homunculi are on the team, and on how the team members work together to produce a corporate output, the account of explanation lacks the resources to distinguish how-possibly from how-actually explanations. One guiding question in this chapter is: How would one have to restrict functional analysis to distinguish how-possibly from how-actually mechanistic explanations?[3]

Second, the account of mechanistic explanation should distinguish mechanism sketches from complete mechanistic models. A *mechanism sketch* is an incomplete model of a mechanism. It characterizes some parts, activities, or features of the mechanism's organization, but it leaves gaps. Sometimes gaps are marked in visual diagrams by black boxes or question marks. More problematically, sometimes they are masked by *filler terms* that give the illusion that the explanation is complete when it is not. A list of common filler terms in neuroscience is shown in Table 4.1. Terms such as "activate," "inhibit," "encode," "cause," "produce," "process," and "represent" are often used to indicate a kind of activity in a mechanism without providing any detail about exactly what activity fills that role. Black boxes, question marks, and acknowledged filler terms are innocuous 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).[4] In contrast, filler terms are barriers to progress when they veil failures of understanding. If the term "encode" is used to

Table 4.1. Common filler terms in neuroscience

Activate	Generate	Process
Cause	Influence	Recognize
Control	Inform	Represent
Encode	Inhibit	Regulate
Excite	Modulate	Store
Filter		

[2] Both the distinctions among how-possibly, how-plausibly, and how-actually descriptions and between a schema and a sketch are introduced in Machamer et al. (2000). (The term "how-possibly explanation" is used in Brandon 1990).

[3] I do not claim that all explanations are mechanistic explanations.

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

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 black-box models as "diagram makers" or "boxologists." Between sketches and complete descriptions lies a continuum of *mechanism schemata* whose working is only partially understood. A second guiding question in this chapter is: how would one have to restrict the systems tradition to distinguish sketches, schemata, and complete descriptions of mechanisms?

Progress in building mechanistic explanations involves movement along both the possibly-plausibly-actually axis and along the sketch-schema-mechanism axis. I now describe how neuroscientists made such progress in the discovery of the mechanism of the action potential.

3. Explaining the Action Potential

The history of the discovery of the mechanism of the action potential serves three purposes in this chapter. First, it provides an example of a successful mechanistic explanation. Second, it illustrates the distinctions that I have just introduced. And, finally, it illustrates many of the norms implicit in the practice of constructing constitutive mechanistic explanations.

I begin where I leave off in Chapter 2. In 1952, Hodgkin and Huxley constructed a mathematical model of the action potential, which they characterize in terms of a list of features (a)–(h) (see Chapter 2, Section 6). They began their project with a background sketch of a mechanism. They knew some of its entities and activities. They knew that action potentials are produced by the movement of ions across a lipid membrane. They knew that action potentials are produced by depolarizing the cell body (that is, by making V_m greater than V_{rest}).[5] And they knew that the shape of the action potential, as described in (a)–(h), could possibly be produced by the voltage-dependent *activation* and *inactivation* of membrane conductance for specific ions, as represented by the variables n, m, and h in their total current equation. Hodgkin and Huxley did not engage in boxology. They used the more informative representational conventions for diagramming electrical circuits shown in Figure 4.1. Left to right, in parallel, are a capacitor and

[5] Here I neglect the possibility of spontaneous action potentials, resulting from the stochastic opening of even a few Na^+ channels.

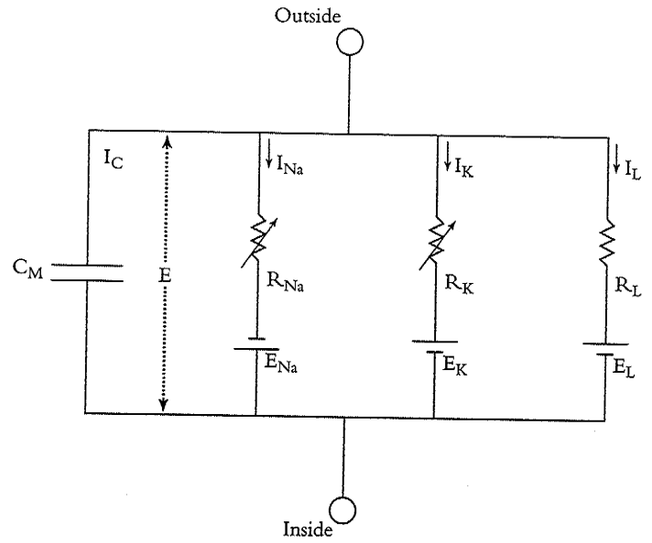

Figure 4.1. The equivalent circuit model of the neuronal membrane*
* I is current, R is resistance, E is the equilibrium potential, and C is capacitance

three pathways for the component currents in the total current equation. Each of the three pathways contains a battery (representing the equilibrium potential for the ion) and a resistor (the inverse of conductance for the ion) in series. The HH model shows that the coordinated changes in the resistances to Na^+ and K^+ currents could account for items a–h.

The primary reason for calling this background mechanistic model a *sketch* is that "activation" and "inactivation" are filler terms. Hodgkin and Huxley (1952) consider some ways to complete these filler terms. They consider the possibility that ions are conveyed across the membrane by active transport. They suggest that perhaps a number of "activation" particles could weaken the integrity of the membrane. They hint at a biological interpretation of their model according to which activation and inactivation particles move around in the membrane and somehow change the membrane's resistance. They admit, however, that they have no evidence favoring their model over other possible models. This admission spurred research on the biophysics of the membrane and the search for ion channels. Nonetheless, well into the 1970s most neuroscientists regarded talk of ion-specific channels as mere metaphor at best and boxology at worst.

C. M. Armstrong (1981) and Bertil Hille (1992) among others elevated talk of ion-specific channels above the status of filler terms. On Hille's model, which is now textbook neuroscience, the conductance changes in action potentials are explained by the temporally coordinated opening and closing of transmembrane channels. Action potentials are generated in the axon hillock, a region at the interface of the cell body and the axon, the "sending" end of a neuron. The hillock is rich in Na^+ channels, and depolarizing the cell body opens these voltage-sensitive Na^+ channels. The resulting increase in membrane conductance (as represented by the dotted line in Figure 4.2) allows Na^+ ions to diffuse from the Na^+-rich extracellular fluid into the relatively Na^+-poor intracellular fluid. This inward Na^+ current is balanced at low values by the effects of depolarization on outward K^+ and leakage currents, the latter of which I ignore for the moment. Above a threshold depolarization, the high voltage sensitivity and rapid activation of the Na^+ channel overwhelms these balancing currents. The flood of Na^+ drives the voltage of the cell towards the Na^+ equilibrium potential (E_{Na}; roughly +55 mV), where the forces of diffusion and voltage balance. This flood accounts for the rapid rising phase of the action potential.

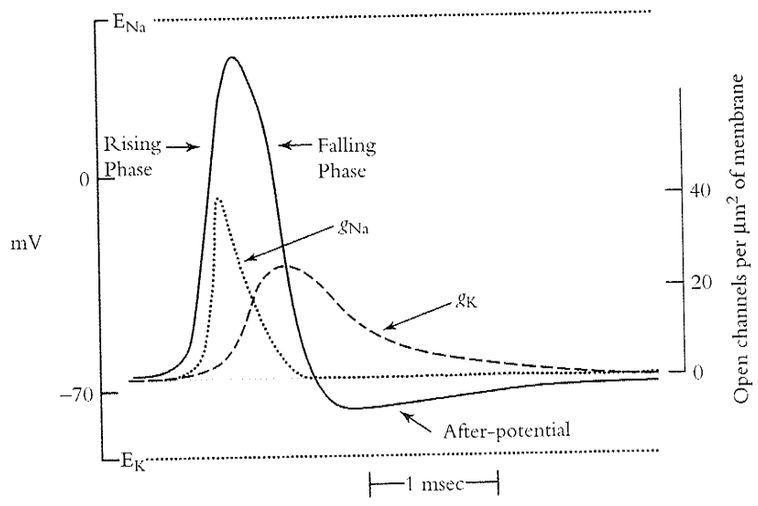

Figure 4.2. The action potential superimposed on a graph of changes in the membrane's conductance for Na^+ and K^+

Depolarizing the membrane has two consequences that account for the declining phase of the action potential. The first is inactivation of Na^+ channels, which slows and eventually stops the ascent of V_m towards E_{Na}. The second is activation of voltage-sensitive K^+ channels, which increases the K^+ conductance of the membrane (as indicated by the dashed line in Figure 4.2) and allows K^+ to diffuse from the K^+-rich intracellular fluid into the relatively K^+-poor extracellular fluid. The diffusion of K^+ out of the cell drives the membrane potential back down towards the K^+ equilibrium potential (E_K; roughly −75 mV) and even below the resting potential of the membrane.

Thus begins the after-potential phase of the action potential, characterized by both hyperpolarization of the membrane (that is, V_m is lower than V_{rest}) and reduced excitability of the neuron. The membrane hyperpolarizes after the action potential because K^+ channels are slow to return to their resting closed state (they are sometimes called "delayed rectifiers" for this reason). The K^+ current tugs V_m away from V_{rest} and towards E_K. This hyperpolarization makes the neuron less excitable, because a larger depolarization is required to move V_m to the threshold for an action potential. This refractory effect is reinforced by the residual inactivation of Na^+ channels, which temporarily prevents them from conducting Na^+ ions. The above is the intermediate elaboration of the action potential mechanism as it appeared when talk of channels gained acceptance through the 1970s and 1980s.

Still, Armstrong and Hille's intermediate elaboration remains a sketch. It fills in some of the details. For example, talk of channels replaces less precise talk about activation and inactivation of conductances, and the focus on channels eliminates speculation about active transport across the membrane. But filler terms remain. In particular, questions remain about how *channels* "activate" and "inactivate." To illustrate how these filler terms were replaced and how how-possibly models gave way to how-actually mechanisms, I focus specifically on how rising membrane voltage can activate and, at higher voltages, inactivate, the Na^+ channel. Hille started work on these mechanisms by conjecturing the set of how-possibly models shown in Figure 4.3. These models of the mechanism have different parts, with different activities, organized in different ways. There are swinging gates, sliding gates, free-floating blockers, tethered balls and chains, rotating cylinders, and assembling components. Hille intended these as merely how-possibly models because he had no idea whether channels would turn

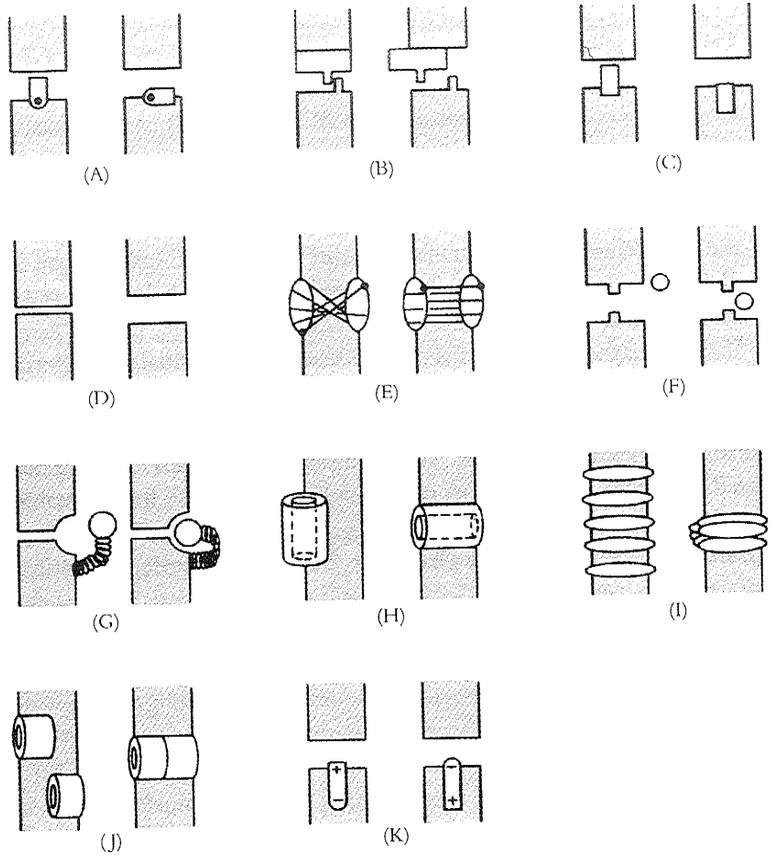

Figure 4.3. Hille's how-possibly mechanisms for gating channels
Source: Reprinted with permission from Hille (1992: 479)

out to have parts of the requisite sort, or whether the parts could act as the model requires, or whether their activities were organized in the way that the model suggests. Hille (1992) rules out many of these how-possibly mechanisms in the face of known constraints and plausibility arguments, leaving only A, B, and C as contenders to account for activation. None of these, however, anticipates the model that subsequently emerged from several independent lines of investigation.

Clues about the Na^+ channel came from sequencing the channel protein, reconstructing its three-dimensional structure, identifying its hydrophilic

and hydrophobic regions (hydrophilic regions are more likely to be inside the membrane, and hydrophobic regions are more likely to be outside) and recording the behavior of individual Na^+ channels under a range of electrical, pharmacological, and genetic interventions.[6] These studies show that the Na^+ channel consists of four subunits, each of which is composed of six membrane-spanning regions (see Figure 4.4). One membrane-spanning region, known as the S4 region, is arranged such that every third amino acid residue is either arginine or lysine. This ordering produces a helical structure, known as an α-helix, with evenly spaced positive charges (see Figure 4.5). At V_{rest}, a positive extracellular potential holds the α-helix in place. Weakening that potential, which happens when the cell is depolarized, allows the helix to rotate out toward the extracellular side (carrying a "gating charge" as positively charged amino acids move outward). This rotation, which occurs in each of the Na^+ channel's subunits, destabilizes the balance of forces holding the channel in its closed state and bends the pore-lining S6 region in such a way as to open a channel through the membrane. Another consequence of these conformation changes is that the pore through the channel is lined with hairpin turn structures, the charge distribution along which accounts for the channel's selectivity to Na^+. Part of the evidence for these conclusions comes from experiments involving site-specific mutagenesis: point mutations induced in the α-helix prevent the channel from opening, and mutations to the hairpin turn regions alter the channel's ion selectivity. The α-helix and the hairpin turn are thus parts of a more complete how-actually model of the rising phase of the action potential.

How does the Na^+ channel close? The currently accepted hypothesis invokes the ball-and-chain model shown in (G) of Hille's diagram. One of the protein subunits composing the channel is thought to contain a long protein strand on the intracellular side of the membrane that terminates in a small "ball" of protein. As V_m reaches a threshold value, this proteinaceous ball and chain swings into the channel, blocking the flow of Na^+ ions. Evidence for this hypothesis includes the fact that removing the ball and chain, either with site-specific mutations or with a pharmacological agent, eliminates inactivation entirely.

[6] For a more detailed discussion, see Catterall (2000) and Hille (1992).

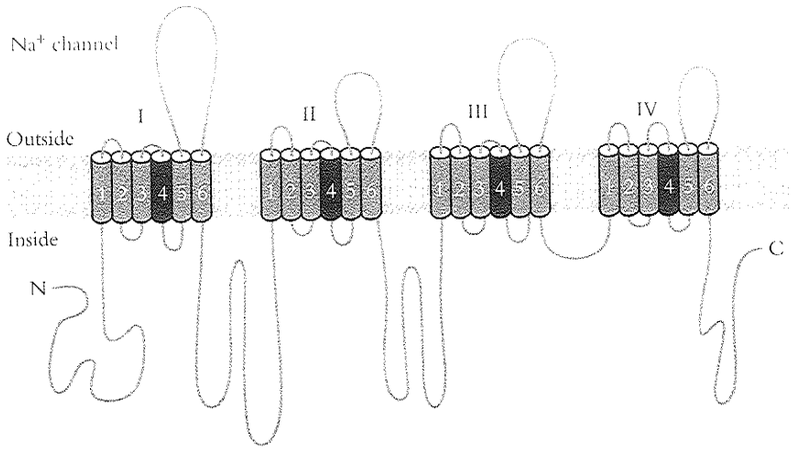

Figure 4.4. Transmembrane regions of the Na^+ channel
Source: Adapted from Hall (1992: 109)

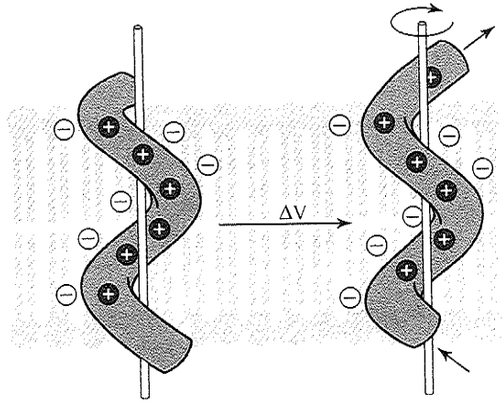

Figure 4.5. A plausible mechanism for activating Na^+ channels*
* An α-helix with regularly spaced positive charges rotates outward
Source: Redrawn from Hall (1992: 112)

There is a great deal more to be said about the mechanisms of Na^+ channel activation and inactivation. Recent studies of the Shaker K^+ using X-ray crystallography and electron microscopy provide detailed accounts of the internal structure of voltage-sensitive K^+ channels

and are leading researchers both to recognize multiple voltage-sensitive regions in channels and to rethink how these voltage-sensitive components work. (For a recent review, see Swartz 2004). The above rotating-helix model of voltage-sensing in the Na^+ channel now has competitors, although it is not currently clear which, if any, is the how-actually model (see Sands, Grottesi, and Sansom 2005). On one model, the S4 region does not move substantially across the membrane but rather causes other parts of the channel to do so, thereby accounting for the gating charge. According to another model, parts of the S4 and S3 regions form paddle-like structures on the external surface of the channel that translocate en masse during voltage changes, thereby opening the channel. According to a third model, two segments within the S4 region twist relative to one another, exposing a channel through the membrane. Although these details are exciting, and although they illuminate the structure of the Na^+ channel, the above textbook sketch is sufficient to reveal the relevant features of mechanistic explanation that I focus on in the remainder of this chapter.

This textbook sketch of one component of the mechanism for the action potential calls attention to three aspects of mechanistic explanation. First, mechanistic explanations are framed by the *explanandum phenomenon* (represented at the top of Figure 4.6), in this case, the action potential as partially described by Hodgkin and Huxley's items (a)–(h). Second, the explanation is *constitutive*; the action potential is explained by reference to component parts of the action potential mechanism. There are component

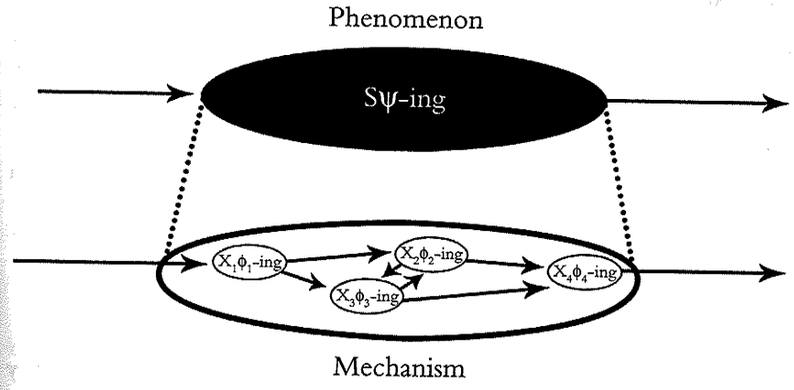

Figure 4.6. A phenomenon (top) and its mechanism (bottom)

entities (the parts), such as ions, ion channels, α-helices, and protein chains, and there are component *activities*, such as diffusion and changes in conformation. The circles and arrows at the bottom of Figure 4.6 represent the mechanism's entities and activities. Third, these entities and activities are *organized* together such that they jointly exhibit the phenomenon to be explained. It matters, for example, that the α-helix contains evenly spaced positive charges, and that the ball at the end of the chain is large enough to block the channel, and that the thresholds for activation and inactivation of the channels are such as to explain the temporal features of the conductance changes. In short, mechanistic models describe how constituent entities and activities are organized to exhibit a phenomenon (compare Bechtel and Richardson 1993; Glennan 1996; Kauffman 1971; Machamer et al. 2000).

Cummins's account of functional analysis can be grafted onto the abstract diagram in Figure 4.6. S's ψ-ing represents the *explanandum phenomenon*, and the circles and arrows represent the analyzing capacities $\{\phi_1, \phi_2, \ldots, \phi_n\}$. This is why functional analysis provides an appropriate starting place for constructing an adequate account of mechanistic explanation. However, Cummins intends his account of explanation to be more general than mechanistic explanation, including in addition what he calls "interpretive explanations," and his effort to develop such a general account prevents him from supplying the kind of detail required for an adequate account of mechanistic explanation. Cummins's commitment to functionalism also leads him to a view of explanation that is abstracted away from the details of the mechanism that realize the functions. This abstraction makes Cummins's account inappropriate as an account of specifically *mechanistic* explanation. In what follows, I show how the basic structure of Cummins's account would have to be elaborated and transformed to provide a normatively adequate account of mechanistic explanations—that is, an account that can distinguish how-possibly from how-actually models, and sketches from complete mechanistic models. Such an account rivals reduction as a normative account of constitutive explanation.

4. The *Explanandum Phenomenon*

The core normative requirement on mechanistic explanations is that they must fully account for the *explanandum phenomenon*. As Kauffman (1971) and Glennan (1996, 2002) argue, mechanisms are always mechanisms *of* a given phenomenon. The mechanism of protein synthesis synthesizes proteins. The mechanism of the action potential generates action potentials. The boundaries of mechanisms—what is in the mechanism and what is not—are fixed by reference to the phenomenon that the mechanism explains. Consider some ways that a description of the phenomenon can fail. These failures provide clues to the standards of success.

One way that the search for mechanistic explanations can fail is by trying to explain a fictional phenomenon. Prior to Galvani's eighteenth-century work on animal electricity, natural philosophers entertained a number of hypotheses about how nerves work (see Pera 1992). Descartes and Borelli believed that nerves are hollow conduits for the flow of animal spirits, and that they activate muscles by inflating them. Starting with this idea, one would be led to search for mechanisms that explain, for example, how nerves shunt the flow of animal spirits into this nerve or that, how they activate muscles, and how light or auditory stimuli impact upon this hydraulic machine. David Hartley, the self-proclaimed Newton of the Mind, believed that neurons work by vibrating. He sought to understand how such vibrations could be distinguished from one another, how they could be stored in the "medullary substance," and how the occurrence of one vibration might cause another to be produced (as demanded by his associationist view of memory). Just as the CL model requires that the *explanandum* sentence should be true, the mechanistic model requires that the *explanandum phenomenon* should exist.

Slightly less obvious are the diverse *taxonomic* errors that one might make in characterizing the phenomenon. If the goal is to provide a mechanistic explanation, the phenomena should be delimited in such a way that they correspond to underlying mechanisms.[7] One kind of taxonomic error is a *lumping error*, which involves assuming that several distinct phenomena are actually one. Cognitive neuroscientists of memory, for example, argue that they have made progress on this front. Daniel Schacter writes that: "We have now come to believe that memory is not a single or unitary faculty

[7] One need not enter the process of discovery with the right taxonomy of phenomena. As Churchland and Sejnowski (1992) illustrate, neuroscientists' understanding of a phenomenon often "co-evolves" with their understanding of underlying mechanisms. Bechtel and Richardson (1993) argue that this co-evolution frequently involves "reconstituting the phenomenon" in the process of searching for mechanisms.

of the mind, as was long assumed. Instead, it is composed of a variety of distinct and dissociable processes and systems. Each system depends on a particular constellation of networks in the brain that involve different neural structures, each of which plays a highly specialized role within the system" (Schacter 1996: 5).[8]

Conversely, one might commit a *splitting error*, which involves incorrectly assuming that one phenomenon is many. For example, it was once assumed that rusting, burning, and breathing are different phenomena with different mechanisms rather than different expressions of a common oxidation mechanism. Taxonomic errors are not always confined to single phenomena, but sometimes infect entire taxonomies. Franz Joseph Gall (1810–19), for example, believed that philosophers were wrong to explain the mind in terms of such mere abstractions as action, memory, perception, cogitation, and will. Gall's system, in contrast, was tailored to identify the set of talents that might vary from individual to individual. His organological map contains cranial regions dedicated to the instinct to murder, tenderness for one's offspring, mechanical skill, facility with colors and coloring, and the impulse to propagation. Contemporary cognitive scientists have a different taxonomy. They divide the mind into such phenomena as motion detection, working memory, change blindness, and pitch perception. The point of this comparison is that it is possible that an entire taxonomic system could be ill-matched to the mechanistic structures of the brain. If so, the taxonomic system is clearly not suited to the search for mechanistic explanations.[9]

One can also err by underspecifying the phenomenon. What is required to fully characterize the *explanandum phenomenon*?[10] For Cummins, the *explanandum phenomenon* is a capacity or disposition, ψ.

To attribute a disposition [ψ] to an object [S] is to assert that the behavior of [S] is subject to (exhibits or would exhibit) a certain law-like regularity: to say [S] has [ψ] is to say that [S] should manifest [ψ] (shatter, dissolve) were any of a certain range of events to occur ([S] is put into water, [S] is struck sharply). (Cummins 1975, in Sober 1984: 401)[11]

[8] Similar arguments are given by Weiskrantz (1990), Schacter and Tulving (1994), and Squire and Knowlton (1994). Such arguments have been discussed by Churchland and Sejnowski (1989); Bechtel and Richardson (1993).

[9] Note that I have not said that this is ground for eliminating the taxonomy. Mismatch between phenomena and the mechanistic structure of the world need not carry eliminativist implications.

[10] Some workers in the systems tradition assume or stipulate that all *explanandum phenomena* have been selected by evolution by natural selection (Lycan 1987, in Lycan 1999: 52–3; Schouten and Looren de Jong 1998: 242–5) or that the phenomena are otherwise adaptive (that is, the phenomenon is how something behaves when it is behaving properly; see Bechtel 1986; Mundale and Bechtel 1996: 485). In the philosophy of biology, Cummins is best known for his attacks on Wright's (1973) adaptive view of functions. I side with Cummins. Neuroscientific explanations often focus on malfunctions, disease states, laboratory phenomena, pharmaceutical contrivances, and industrial and military applications (for example, how the vestibular system works in zero-gravity). There also seems no reason to presuppose that all of the functions currently operating in organisms have selective histories. Traits can become entrenched through genetic drift and exaptation. The considerations below also provide reasons for concluding that adaptive functional characterizations of the phenomenon omit much of the crucial information for distinguishing how-possibly from how-actually models. No doubt, some of the features of the brain have straightforward adaptive etiologies, but I do not want to presuppose for present purposes that all of them do. Either way, one still needs the more limited sense of role-functions, activities that make some crucial contribution to the behavior of a containing system (Cummins 1975; Craver 2001).

[11] I have changed the variables for consistency.

[12] Cummins (2000) has explicitly abandoned this connection between functional analysis and laws (more about which in Section 6).

Several years later, Cummins reiterates: "A capacity is specified by giving a special law linking precipitating conditions to manifestations—that is, by specifying input-output conditions" (Cummins 1983: 53). This terse characterization of the phenomenon downplays the wealth of detail that can be used to distinguish how-possibly from how-actually explanations and to distinguish sketches from complete descriptions of mechanisms.[12] Consider some dimensions along which this basic characterization might be elaborated.

Phenomena are typically *multifaceted*. Action potentials are complex phenomena, when compared to shattering and dissolving. 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 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. Consider, for example, how different values of the peak magnitude of the action potential demand (and exclude) different mechanistic explanations. In 1902, Julius Bernstein hypothesized that nerve impulses might be produced by a sudden breakdown in membrane resistance. If so, the action potential should peak at a value no higher than 0 mV. And so it was widely believed until the 1930s, when Kenneth Cole and Howard Curtis (1939) confirmed that the membrane resistance drops by roughly two orders of magnitude during

the action potential. However, Hodgkin and Huxley (1939) demonstrated that the action potential overshoots 0 mV, peaking at +40 to +50 mV. Curtis and Cole (1940) later redid these experiments with a different kind of electrode and found much higher peak magnitudes. They published one of their most dramatic examples of the overshoot, which peaked at +110 mV. Because the Na^+ equilibrium potential is roughly +55 mV, Hodgkin and Huxley's finding was consistent with the possibility that the rising phase of the action potential is constituted by a breakdown in the resistance to Na^+. Curtis and Cole's finding, on the other hand, could not be explained by that mechanism (or at least not by that mechanism alone). Note that this is only one of multiple aspects of the manifestation of the action potential. If the peak magnitude can be characterized as an "output," then the action potential is characterized by a very large array of input–output relationships, each of which must be satisfied by any explanatory model of the mechanism. The 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 merely a 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 mechanism of the action potential, one should be able to say why they are *not* produced under these conditions. A complete characterization of the phenomenon also requires knowing the phenomenon's *modulating conditions*—that is, knowing how variations in conditions alter the action potential. For example, one wants to know how the action potential changes if one changes the neuron's diameter, or the density of ion channels in a given stretch of membrane, or the extracellular concentration of Na^+. 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 nonstandard. If one connects a squid giant axon (the experimental system in which most of these 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),[13] 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 nonstandard conditions.

A variety of *by-products* or side effects of the phenomenon can also be crucial for sorting how-possibly from how-actually models and sketches from complete mechanistic models. By-products 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. Cummins now recognizes that describing phenomena as capacities is an oversimplification, and that it is often "a matter of some substance" to specify what the *explanandum* is (Cummins 2000: 123–4):

> Given two theories or models of the same capacity, associated incidental effects can be used to distinguish between them. ... Even when two models are not weakly equivalent, they may be on a par empirically, that is, close enough that the differences between them are plausibly attributed to such factors as experimental error, idealization, and the like. Again, incidental effects that may have no great interest as explananda in their own right may serve to distinguish such cases (2000: 124).

As noted above, the activation of Na^+ channels is accompanied by a gating charge, a very slight movement of charges across the membrane. Why

[13] "Normal" and "standard" conditions amount to something like "the way that the mechanism behaves under the conditions that we consider most appropriate for our current explanatory purposes." Sometimes this is assessed in terms of the healthy and fit organism, and normal means something like "behavior consistent with or conducive to overall system health and function." Sometimes it is assessed in terms of evolutionary stories, and so means something like "behavior similar to that which preserved the trait in the population of organisms." Sometimes normalcy is assessed in terms of its utility for an experiment, and so means something like "behavior consistent with or conducive to manipulation and detection with my experimental protocol." There is no need to be more restrictive about this notion. "Normal" and "standard" are defined relative to an implied investigative context.

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 (see Figure 4.5). 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 competing models of voltage sensor mentioned in Section 3 have to, and are designed to, accommodate the gating charge. 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.[14]

In summary, mechanistic explanations can fail because one has tried to explain a fictitious phenomenon, because one has mischaracterized the phenomenon, and because one has characterized the phenomenon to be explained only partially. 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 the multiple features of a phenomenon, of its precipitating, inhibiting, modulating, and nonstandard conditions, and of its by-products, all constrain mechanistic explanations and help to distinguish how-possibly from how-actually explanations. Similarly, mechanism sketches, with large gaps and question marks, can 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 component currents, but the sketch does not explain the conductance changes that regulate the current flow. To characterize the phenomenon correctly and completely is a crucial step in turning a functional analysis into an acceptable mechanistic explanation.

5. Components

Mechanistic explanations are *constitutive* or componential explanations: they explain the behavior of the mechanism as a whole in terms of the organized activities and interactions of its *components*. Components are the entities in a mechanism—what are commonly called "parts." Action potentials are

[14] Such by-products are not functionally relevant but they nonetheless are part of the phenomenon to be explained. This is an additional reason to think that the character of the phenomenon should not be restricted to those features that contributed to survival, or that contribute to current health, etc.

explained by appeal to components such as Na^+ and K^+ channels, ions, and protein chains.

This is a crucial point of contrast between my mechanistic view of constitutive explanation and Cummins's account of functional analysis. Unlike other exponents of the systems tradition, Cummins insists that, "it is important to keep functional analysis and componential analysis conceptually distinct" (Cummins 1983: 29, 2000: 123). He insists on this point because he wants to allow for *non*constitutive analytic explanations—analytic explanations in which both the analyzed and the analyzing capacities (ψ and ϕ, respectively) are capacities of the system as a whole.[15] He gives the following example: if one wants to explain how a cook (S) bakes a cake (ψ), one will appeal to analyzing capacities (ϕ) that are also "cook-level" capacities, such as reading recipes, stirring, and salting to taste. He also discusses John B. Watson's explanation of maze-running in terms of capacities such as stimulus substitution and "the ability to respond in certain simple ways to simple stimuli," which are also properties of the rat as a whole (1975: 761).[16] Cummins makes this allowance to accommodate "interpretive explanations," which appeal to the flow of information or to the manipulation of representations in a system. Indeed, Cummins is not primarily interested in constitutive mechanistic explanations, but rather with forms of psychological explanation that are functional and largely independent of the implementing mechanisms. I agree with Cummins that these two varieties of explanation must be kept distinct, especially in discussions of explanation in neuroscience. Lumping both together under the rubric of functional analysis blurs this distinction. So let us make it explicit that functional analysis and mechanistic explanations are distinct in that in mechanistic explanations, *S's* ψ-ing *is not explained merely by the subcapacities of* ψ, *but by the capacities* $\{\phi_1, \phi_2, \ldots, \phi_n\}$ *of S's component parts* $\{X_1, X_2, \ldots, X_m\}$.

The distinction is crucial because how-actually explanations are often distinguished from how-possibly explanations on the grounds that the latter appeal to component parts that do not exist and because models

[15] Cummins often frames his account of functional analysis without any reference to component parts. For example: "Functional analysis consists in analyzing a disposition into a number of less problematic dispositions such that the programmed manifestation of these analyzing dispositions amounts to a manifestation of the analyzed disposition" (2000: 125).

[16] Thanks to Uljana Feest for calling my attention to this ambiguity.

of mechanisms are often distinguished from sketches on the grounds that the latter contains black boxes or filler terms that cannot be completed with known parts or activities. In some functional explanations (such as interpretive explanations), explanations describe a *program* that *could possibly* produce the phenomenon. Cummins repeatedly emphasizes that, "there is no unique right answer to the question 'Which program does this system execute?'" (1983: 30–43). Further:

> Any way of interpreting the transactions causally mediating the input-output connection as steps in a program for doing [ψ] will, provided it is systematic and not *ad hoc*, make the capacity to [ψ] intelligible. Alternative interpretations, provided they are possible, are not competitors; the availability of one in no way undermines the explanatory force of another. (Cummins 1983: 42; symbol substituted for consistency)

For interpretive functional explanations, then, any set of how-possibly ϕ-ers will suffice so long as they can be strung together in a program that accounts for S's ψ-ing. Not so for mechanistic explanations. If it did suffice, then Hodgkin and Huxley would have counted their equations for the conductance changes as explanations, but as I show in Chapter 2, they insist that their proposed sequence of biological activities, involving activation particles and their motion in the membrane, is only a convenient fiction.

In a more recent paper, Cummins acknowledges that in neuroscientific explanations in particular one cannot be so cavalier about how different psychological capacities are realized in the parts of organisms:

> Neuroscience enters the picture as a source of evidence, arbitrating among competitors, and ultimately, as the source of an account of the biological realization of psychological systems described functionally. (Cummins 2000: 135)

And furthermore:

> a complete theory for a capacity must exhibit the details of the target capacity's realization in the system (or system type) that has it. Functional analysis of a capacity must eventually terminate in dispositions whose realizations are explicable via analysis of the target system. Failing this, we have no reason to suppose we have analyzed the capacity as it is realized in that system. (Cummins 2000: 126)

No neuroscientist would claim that it makes no difference whether action potentials are produced by passive diffusion through Na^+ and K^+ channels or by active transport through some energy-intensive membrane

mechanism. Mechanistic explanation is inherently componential. Box-and-arrow diagrams can depict a program that transforms relevant inputs onto relevant outputs, but if the boxes and arrows do not correspond to component entities and activities, one is providing a redescription of the phenomenon (such as the HH model of conductance changes) or a how-possibly model (such as their working model of conductance changes), not a mechanistic explanation.

Distinguishing good mechanistic explanations from bad requires that one distinguish real components from fictional posits. The most dramatic examples of fictional posits include animal spirits, entelechies, and souls, but fictitious entities can be far more mundane than these. It might have turned out that Bertil Hille's channels did not exist. The movement of charge across the membrane might well have been a matter of active transport (as Hodgkin and Huxley once thought), or degradation of the membrane (as Bernstein suggested), or it might have involved no movement of ions across the membrane at all. Many of the how-possibly mechanisms in Figure 4.3 require parts (and activities) that do not exist.

There is no clear evidential threshold for saying when one is describing real components as opposed to fictional posits, or for detecting when one is pushing one's hypothesis a bit far (as Hille's older colleagues claimed). Nonetheless, the following four criteria are satisfied by real parts and help to distinguish mere how-possibly from how-actually explanations. Real parts have a stable cluster of properties, they are robust, they can be used for intervention, and they are physiologically plausible in a given pragmatic context.

First, the parts should have a *stable cluster of properties* (compare Boyd 1991). Hille's speculative channels were gradually transformed into stock-in-trade entities as it became possible to identify them as proteins, to determine the linear order of their amino acids, to recover their secondary and tertiary structure, to describe their interactions with neurotransmitters and with chemical agonists and antagonists, to characterize their voltage-dependence and rapid inactivation, and so on. Discovering clusters of such properties also allowed researchers to distinguish multiple kinds of channel proteins, selective for different ions, sensitive to different agonists and antagonists, composed of different sequences of amino acids, and the like. As details mounted about the shapes of the channels, their components, their causal interactions, and their subtypes, it became increasingly difficult to dismiss channels as a mere hypothesis being "pushed too far."

Second, and related, the parts should be *robust* (Wimsatt 1981). They should be detectable with a variety of causally and theoretically independent devices. 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 for 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 (Salmon 1984; Culp 1994, 1995; Psillos 1999; Achinstein 2002).

Third, it should be possible to use the part to *intervene* into other parts 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, and to balance current. As I argue in Chapter 3, the ability to manipulate items in this way is crucial evidence for establishing causal and explanatory relationships among the mechanism's components. This criterion is also crucial for distinguishing real components from fictional posits.

Finally, the component should be *physiologically plausible*. It should not exist only under highly contrived laboratory conditions or in otherwise pathological states. Of course, 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 of the phenomenon

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 (black or grey boxes), 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. To adequately describe mechanistic explanation, functional analysis must be restricted to constitutive explanations in which some property or activity of a whole is explained in terms of the properties or activities of its parts. And it should be restricted to cases in which the components in the explanation are not mere fictions but real components in the system.

6. Activities

The systems tradition and Cummins's account usefully shift attention away from etiological explanation and toward the kinds of explanation found in neuroscience and psychology. As Cummins notes:

> The concern to distinguish causal laws from noncausal correlations, to shun uncaused or idle events, and to make provision for independent access to causes and effects are, of course, not the only methodological concerns to manifest themselves in scientific practice and in writings on the scientific method, but they are, perhaps, the most fundamental and pervasive. ... It should become clear shortly, however, that these concerns are simply out of place in the context of property theories and the analytic strategy of explanation. (1983: 14)

Cummins is right to call attention to this philosophical tunnel vision. However, functional analyses are made up of capacities, and mechanisms are partly constituted by activities and interactions. An adequate account of constitutive explanation must address traditional philosophical problems about causation.

Cummins describes capacities as input–output relations that relate precipitating conditions to manifestations. He requires further that there must be "laws *in situ*" relating input to output conditions, that is, "laws that hold of a special kind of system because of its peculiar constitution and organization. The special sciences do not yield general laws of nature, but rather laws governing the special sorts of systems that are their proper objects of study" (2000: 121). Laws *in situ* are what I describe in Chapter 3 as mechanistically fragile generalizations.

Cummins's appeal to laws to analyze capacities raises two issues. The first is the matter of squaring this view with his insistence that laws are not explanatory. Cummins rejects the CL model because "No laws are explanatory in the sense required by the [CL model]. Laws simply tell us what happens; they do not tell us why or how" (2000: 119). If capacities or dispositions are analyzed in terms of special laws, and special laws are not explanatory, then it is hard to make sense of Cummins's claim that "[functional] analysis allows us to explain how the device as a whole exercises the analyzed capacity, for it allows us to see exercises of the analyzed capacity as programmed (that is, organized) exercises of the analyzing capacities" (2000: 125). Cummins needs to distinguish the laws used in CL explanations from the laws underlying the capacities in functional analyses, but he does not articulate the difference.

The second issue is that Cummins needs a way to distinguish bona fide capacities from pseudo-capacities. One can use Cummins-style input–output pairs to describe mere temporal sequences (input crowing roosters, output dawn), effect-to-cause pairs (input refractory period, 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). It will not help to require that the input–output regularity support counterfactuals (as Weber 2005 requires), because not all counterfactual 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. Cummins requires an account of analyzing capacities sufficiently robust to satisfy criteria such as (E1)–(E5).

In Chapter 3, I show that the manipulability account distinguishes laws from accidents while honoring Cummins's accurate assessment that the laws of neuroscience and psychology are mechanistically fragile. The causal relationships in mechanisms are not mere capacities in Cummins's sense; they are relationships that are potentially exploitable for purposes of control.

7. Organization

Mechanistic explanatory texts can begin with a correct and complete characterization of a phenomenon, and with real parts and bona fide

capacities, and still fail to understand how these components are *organized*. Cummins defines organization as something that can be described in a flow chart or a program. But almost anything can be described in a flow chart or a program. Arguments, libraries, time-lines, taxonomic systems, chains of command, legal precedent, and words in a book all can be described with boxes and arrows. Yet some forms of organization are distinctively mechanistic.

The distinctively mechanistic form of organization can be brought out by contrasting mechanisms with mere *aggregates* (Wimsatt 1985, 1997). In an aggregate, the whole is literally the sum of its parts. Suppose that a property or activity (ψ) of the whole (S) is explained (in an ontic sense) by the properties or activities $\{\phi_1, \phi_2, \ldots, \phi_n\}$ of its parts $\{X_1, X_2, \ldots, X_m\}$. The ψ-property of S is an aggregate of the ϕ-properties of X's when:

(W1) ψ is invariant under the rearrangement and intersubstitution of Xs;
(W2) ψ remains qualitatively similar (if quantitative, differing only in value) with the addition or subtraction of Xs;
(W3) ψ remains invariant under the disaggregation and reaggregation of Xs; and
(W4) There are no cooperative or inhibitory interactions between Xs that are relevant to ψ. (Modified from Wimsatt 1997)

Wimsatt's (W1)–(W4) are criteria for diagnosing the importance of organization in a system and are also a set of strategies for discovering a mechanism's organization. Compare the mechanism for generating action potentials (S_1) to a neat glass of gin (S_2) and, likewise, Na^+ channels, K^+ channels, and the membrane (the Xs in S_1) to unit volumes of gin (the Xs in S_2). The mechanism of the action potential (S_1) generates action potentials (ψ_1), and the glass of gin (S_2) has a given volume (ψ_2). The parts (Xs) of the action potential mechanism, such as the Na^+ channels and K^+ channels, cannot be intersubstituted with one another (W1). They are ion specific. Only judicious removal or multiplication of certain parts of the action potential mechanism is compatible with its continued working as a whole (W2). Changing even the spatial relations among the components of the mechanism would, at least in many cases, completely disrupt the behavior of the whole system (W3). Finally, there are cooperative and inhibitory interactions between the components of the action potential mechanism (W4). This is why alterations to the entities and activities of the

action potential mechanism can destroy it. The volume of gin, in contrast, stays the same any way you shake it.

The above description of the mechanism of the action potential displays three varieties of mechanistic organization: active, spatial, and temporal. Different kinds of organization predominate in different mechanisms.

Most fundamentally, the components in the action potential mechanism are *actively organized*. In direct violation of W4 they act and interact with one another in such a way that the ψ-ing of S is more than just a sum of ϕ-properties. In fact, the ϕ-properties of a working mechanism are not just properties; they are the activities of and interactions among the entities in the working mechanism. The different components act in cooperation or competition, and they do so with some components and not with others. It matters which Xs ϕ with which others, and it matters how they interact. This is why the parts of mechanisms often cannot be reorganized randomly (W1), added or subtracted at will (W2), or taken apart and put back together again (W3) without disturbing their corporate ability to ψ.

Active organization also distinguishes mechanistic explanations from what John Haugeland calls "morphological explanations." Haugeland (1998) illustrates the contrast with the transmission of an image along a fiber optic cable. An image projected on one end of the cable is transmitted to the other by an array of bundled fibers. Each fiber is an isolated conduit of light for a given dot in the image, the brightness and color of which is transmitted along the length of the wire. So long as the relative spatial arrangement of the wires in the bundle is the same at each end, the input image is conserved in the output image. In such explanations, mere spatial organization does most of the work. The fibers do not relevantly interact with one another or work together, and they can become hopelessly tangled in the middle of the wire so long as the spatial arrangement remains the same at the end. This relationship is a matter of degree, however, because many morphological explanations also essentially involve interactions among the parts (for example, the shapes of crystals are determined by the shapes of the molecules, their spatial arrangement, and their packing). Mechanisms, in contrast, are not mere static or spatial patterns of relations, but rather patterns of allowance, generation, prevention, production, and stimulation. There are no mechanisms without active organization, and no mechanistic explanation is complete or correct if it does not capture correctly the mechanism's active organization.

Finally, active organization distinguishes mechanistic models from taxonomic schemes or temporal sequence displays. The periodic table organizes the elements in terms of their underlying atomic structures by exhibiting, for example, some as noble gases and others as radioactive isotopes. Although there are mechanistic models for explaining why certain elements are possible or impossible, and although these allow one to predict the existence of hitherto unobserved elements, mere taxonomic ordering is not a mechanism. Similarly, the Linnaean taxonomic system, which sorts organisms on the basis of their phenotypic traits, is not a mechanistic explanation: the components in this system do not *do anything* that contributes to a behavior of all of the parts taken together. The same can be said of purely sequential theories, such as those describing developmental stages of an organism or the life cycle of a cell. Purely sequential models describe time-slices of a four-dimensional object and show how its parts are arranged at different stages. However, such models do not show how one stage arises from its predecessor, or how the configuration of parts at one stage produces the configuration at the next.

Active organization in mechanisms is sustained by the *spatial* and *temporal organization* of the component parts. The same entities and activities joined in different spatial and temporal organization often yield different mechanisms. The spatial organization of a mechanism includes, for example, the sizes, shapes, structures, locations, orientations, directions, connections, and compartments of its components. Several kinds of spatial organization are crucial for understanding the mechanism of the action potential. It matters, for example, that the ion channels have appropriate *sizes* to allow the flow of ions, that they are long enough to traverse the membrane, and that they are small enough to fit in a small patch of membrane. It matters that the components have appropriate *orientations*, for example, that the ball and chain is inside the membrane and that the α-helix is roughly perpendicular to the membrane. The conformations (or *shapes*) of the ion channels under different membrane voltages allow them to act as channels and to gate the flow of ions appropriately. Furthermore, it matters that large numbers of Na^+ channels are *located* in the axon hillock, and that the α-helix is in the S4 region of the protein.[17] In many mechanisms, it

[17] Bechtel and Richardson (1993) emphasize the importance of localization in discovering mechanistic explanations.

matters that the different components are in *contact* with one another: it matters that the S4 region is *connected* to the rest of the channel or that the ball and chain comes into contact with the walls of the channel. This is not an exhaustive list of spatial forms of organization found in mechanisms. Furthermore, one should not expect each form to be equally represented in all mechanisms. Some mechanisms, such as biochemical cascades in the cytoplasm, depend less on the precise location of the activities than on the structures of the entities involved and the temporal arrangement of their activities. But in many cases—and the action potential is an excellent example—spatial organization provides the structure by virtue of which mechanisms work. Getting the relevant aspects of the spatial organization right is part of developing a good and complete mechanistic explanation.

The third aspect of mechanistic organization, is temporal. The order, rate, and duration of successive component activities are crucial for the action potential. There is a sequence of stages from beginning to end, and it is not possible to change their order without interfering with how the mechanism works (or making it a different mechanism entirely). The activation and inactivation of the Na^+ and K^+ channels are appropriately timed so that the action potential rises, falls, and exhibits its characteristic refractory period. The rates at which the channels open and close, and the duration over which they are open or inactivated, are similarly crucial for the overall shape of the action potential. One much-noted problem with programs in classical artificial intelligence is that they often ignore real-world temporal constraints on processing. One might be able to simulate object recognition in LISP, but such a model is unlikely to work as fast as the visual system. Good mechanistic explanations incorporate temporal constraints.

In Cummins's account, the notion of organization is underspecified, requiring only that it be possible to describe the system with a box-and-arrow diagram. He requires this abstraction in order to accommodate interpretive functional explanations (such as Watson's explanation for how a mouse runs a maze). Mechanistic explanations, however, are embodied. They are anchored in components, and those components occupy space and take time to act. A description of a mechanism is not merely a summation of parts or capacities; it is a description of how they work together. That description involves—in addition to a list of component entities $\{X_1, X_2, \ldots, X_m\}$ and activities $\{\phi_1, \phi_2, \ldots, \phi_n\}$—an account of how they are organized together actively, spatially, and temporally in S's ψ-ing.

As the discovery of the mechanism of the action potential illustrates, neuroscientists distinguish how-possibly from how-actually models by adding such constraints on organization (see Craver and Darden 2001). In the final chapter of this book, I show how the mosaic unity of neuroscience is achieved when different fields, with different techniques and theoretical vocabularies, place different constraints on the same mechanism.

8. Constitutive Relevance

To recap the basic features of my account thus far: The *explanandum* of a mechanistic explanation is a phenomenon, typically some behavior of a mechanism as a whole.[18] The central criterion of adequacy for a mechanistic explanation is that it should account for the multiple features of the phenomenon, including its precipitating conditions, manifestations, inhibiting conditions, modulating conditions, and nonstandard conditions. The explanans is a mechanism. The model of a mechanism does not describe capacities of the mechanism as a whole; it describes the activities of the mechanism's components. How-possibly models can be composed of fictional components, but how-actually models describe real components that have multiple properties, that are detectable with multiple techniques, that are utilizable for the purposes of intervention, and that are physiologically relevant. The model of the mechanism also describes the causal relations (activities) that compose the mechanism. These are not mere input–output relationships or *laws in situ* but relationships of manipulability as described in Chapter 3. Finally, mechanistic explanatory texts do more than exhibit box-and-arrow diagrams; they reveal the active, spatial, and temporal organization of a mechanism. These restrictions make significant progress in defining mechanistic explanation, in distinguishing it from other kinds of explanation, and in distinguishing good explanations from bad.

However, this model of mechanistic explanation is not yet complete. There are two reasons: first, I have not said what it means for a model of a mechanism to "account for" the phenomenon. According to the CL model—the model of explanation at the core of classical reduction—one accounts for the phenomenon by showing that its diverse features *are to be*

[18] In an explanatory text, the *explanandum* is a description of the phenomenon and the *explanans* is a description, or schema, of a mechanism.

expected on the basis of the description of the mechanism; this means that one can infer the features of the phenomenon from a complete specification of the mechanism, the initial conditions, the background conditions, and the relevant transtheoretic identities. In Chapter 2, I argue that this is an inadequate view of the nature of explanation on the grounds that it is impossible for most explanations in neuroscience and that it is too weak to distinguish explanatory derivations from, for example, mathematical models that merely save the phenomena. However, I have not yet provided an alternative vision of how a mechanism accounts for, and so explains, the phenomenon.

Second, I have not provided an account of *constitutive* explanatory relevance.[19] That is, I have not said when a part of S is a component in the mechanism of S's ψ-ing. Not all parts are components. Consider again the difference between mechanisms and machines. Machines contain many parts that are not in any mechanism. The hubcaps, mud-flaps, and the windshield are all parts of the automobile, but they are not part of the mechanism that makes it run. They are not *relevant* parts of that mechanism. Good mechanistic explanatory texts describe all of the relevant components and their interactions, and they include none of the irrelevant components and interactions. The failure to address constitutive relevance is a major lacuna not just in Cummins's model of explanation, but also in the systems tradition generally, in recent discussions of mechanistic explanation (including my own), and, in fact, in all discussions of "mircroreduction" in the philosophy of biology and the philosophy of mind. Considerable philosophical effort has been expended on the topic of etiological (that is, causal) relevance, but almost none has been dedicated to the problem of constitutive relevance.

In Section 8.1, I show that any adequate account of mechanisms must supply an account of constitutive relevance. I build my positive account by considering the experimental strategies that neuroscientists use to test whether a given entity, activity, property, or organizational feature is relevant to the behavior of the mechanism as a whole and by considering some well-known ways that these strategies can fail (8.2). The account of constitutive relevance should block these failures in much the same way that

[19] Again, I am using this term as Salmon uses it, that is, to refer to an underlying mechanism. His goal is to specify the sense in which a component is relevant to, and so is part of the explanation for, the phenomenon.

the manipulationist account of etiological relevance in Chapter 3 blocks the kinds of failures expressed in (E1)–(E4).[20] I then offer a sufficient condition for interlevel relevance: the *mutual manipulability* account. According to that account, a part is a component in a mechanism if one can change the behavior of the mechanism as a whole by intervening to change the component *and* one can change the behavior of the component by intervening to change the behavior of the mechanism as a whole.

8.1 Relevance and the boundaries of mechanisms

One cannot delimit the boundaries of mechanisms—that is, determine what is in the mechanism and what is not—without an account of constitutive relevance. To see that this is the case, consider the shortcomings of some efforts to delimit the boundaries of mechanisms without appealing to explanatory relevance.

One might, for example, equate the boundaries of mechanisms with compartmental boundaries. Some mechanisms are entirely contained within physical compartments, such as a nucleus, or a cell membrane, or skin. Transcription (typically) happens within the nucleus, and translation occurs in the cytoplasm. Detection of plasma ion concentrations happens within the circumventricular organs and outside of the blood–brain barrier. However, mechanisms frequently transgress compartmental boundaries. The mechanism of the action potential relies crucially on the fact that some components of the mechanism are inside the membrane and some are outside. The membrane allows the intracellular and extracellular concentrations of ions to be different, allows a diffusion gradient to be set up, and allows for a separation of charge. Likewise, many cognitive mechanisms draw upon resources outside of the brain and outside of the body to such an extent that it may not be fruitful to see the skin, or the surface of the CNS, as a useful boundary (as Haugeland 1998; Wilson 1995, 2004; Clark 1997; and Clark and Chalmers 1998 emphasize). Examples such as this are commonplace.[21]

Cartesian mechanists faced this challenge as well. If the extended world is devoid of goals and purposes, composed only of corpuscles operating

[20] The failure of (E5) is not germane in the case of constitutive explanations.

[21] von Eckardt and Poland (2005) criticize my view of mechanisms in Craver (2001) for my failure to accommodate outward- and upward-looking explanations. Craver (2001) is motivated in part by the desire to accommodate such explanations.

blindly by motion and contact, what principles could possibly define the unity of a machine, organ, or organism? Descartes at times favors principles of spatial organization: the parts are within a spatial boundary, they move together, and they can be transported together from one place to another while maintaining fixed relative positions with the other components (see Des Chene 2001). Others (such as Salamone De Caus) appeal to contact among the parts. Few contemporary scientists hold to the idea that all causal interactions require contact among components, but even granting this possibility, there are several counterexamples to each of these suggestions. I have already noted that mechanisms frequently defy tidy physical boundaries (although every mechanism can, trivially, be circumscribed). Parts of mechanisms often move in separate directions (as any multiple-pulley system illustrates). Some mechanisms are more ephemeral than others;[22] they work only as components happen to come into the appropriate spatial arrangement. For example, in many biochemical cascades, the relevant reactions could happen anywhere in the cytoplasm. Such mechanisms lack stable spatial relations; they cannot be picked up and carried from one place to the next.

Some members of the systems tradition define the boundaries of mechanisms by appeal to the *intensity of interaction* among its components. Herbert Simon (1969) takes this approach in his discussion of "near complete decomposability." Systems, for Simon, are sets of state variables and their interactions. A system can be decomposed into distinct (that is, bounded) subsystems by comparing the relative strengths of interactions among the variables in the system as a whole. Variables are clustered into subsystems when their interactions with one another are stronger than are their interactions with variables outside of that set. Wimsatt (1976b) adopts the same view, but notes that the threshold of strength required for inclusion in a mechanism depends upon one's pragmatic interests. If one requires exacting control or exceptionally precise predictions, then even weak interactions must be included in the model. For other purposes, one might be willing to tolerate error or imprecision and so can neglect weak interactions. John Haugeland develops a third variant. He describes systems as "relatively independent and self-contained composites of components interacting at interfaces" (1998: 215), where to be relatively independent and

[22] I borrow this term from Stuart Glennan (personal communication).

self-contained is to interact more often and more intimately with items inside the interfaces than with those outside (215). Intimacy, in turn, "means something like how 'tightly' things are coupled, or even how 'closely knit' they are" (215). Grush (2003) has elaborated Haugeland's position (for the purposes of criticizing it), as distinguishing mechanisms from their environment by the *bandwidth* of their interaction. By bandwidth, Grush means the number of state variables describing a component that appear in the equations describing the evolution of the other state variables. The narrower the bandwidth, the more appropriate it is to identify the interaction as an interface. Grush's elaboration of Haugeland's view is the most sophisticated interactionist account of the boundaries of mechanisms.[23]

The bandwidth criterion, it would seem, needs to be supplemented by the kinds of pragmatic considerations that Wimsatt discusses in order to specify a threshold of bandwidth below which the components are separate and above which the components are not separate. The location of this threshold is likely to depend upon one's error-tolerance and one's purposes. Leaving this issue aside, however, neither strength nor bandwidth criteria suffice to pick out the right boundaries. First, they do not readily distinguish components from *background conditions*. The beating of my heart and my ability to read are strongly connected to one another, on any notion of causal strength. If my heart were to stop beating for any stretch of time, or if it were to speed up dramatically, my ability to read would quickly decay. The action potential shares high-bandwidth interfaces with protein synthesis mechanisms and glucose metabolism, for example, because ion channels are constantly recycled in the membrane and because the ion pump that establishes the V_{rest} is energetically demanding. Although the distinction between a mechanism and a background condition is likely to be vague, it is nonetheless a common distinction and one that a view of constitutive relevance can help to sort out.

There is also the problem of *sterile effects*. Components of mechanisms have many effects that are irrelevant to the behavior of the mechanism.

[23] Grush (2003) develops his own view, according to which distinct mechanisms are plug-and-play components that can readily be ejected from and plugged into a systemic context. This works well for the case that Grush considers (the importance of the skin as a boundary for cognition), but it works less well for biochemical cascades and physiological systems, which are often spatially quite distributed and so tightly interwoven into their systematic context that it is very difficult to see them as self-contained in this way. Furthermore, whether a mechanism is truly plug-and-play depends on whether it contains all of the relevant components.

Such effects might provide evidence for or against a given how-possibly mechanism, but they are not part of the mechanism itself. Action potentials affect a host of other processes in cells, such as protein synthesis, metabolism, membrane turnover, and packaging of neurotransmitters. Na^+ channels exert attractive and repulsive effects on ions and other particles in the cytoplasm and in the extracellular space, they deform the membrane, and so on. But these activities are sterile in the mechanism: they either produce no changes in the other components of the mechanism, or the changes they do produce make no difference to action potentials. In contrast, the rotation of the α-helix, the movement of the ball and chain, and the diffusion of ions are all tightly coupled in a way that does make a difference to action potentials.[24]

The boundaries of mechanisms, it appears, cannot be defined by strength of interaction or bandwidth alone. The spatial and interactive boundaries of mechanisms depend on the epistemologically prior delineation of *relevance boundaries*. Spatial boundaries are those that circumscribe all the relevant entities and activities. Temporal boundaries are those that include all the relevant activities. An account of constitutive mechanistic explanation must include an account of constitutive relevance. The causal-mechanical alternative to derivation as a regulative ideal is that the mechanism should describe all of the components, activities, and organizational features that are *relevant* to the *explanandum phenomenon*.

8.2 *Interlevel experiments and constitutive relevance*

The norms of constitutive relevance are implicit in the experimental strategies that neuroscientists use to test claims about componency and in the rules by which neuroscientists evaluate applications of those strategies (Bechtel forthcoming; Bechtel and Richardson 1993; Craver 2002b). Neuroscientists use these experiments to establish which parts are components in a mechanism and which are not, that is, to distinguish relevant components from mere constitutive correlates (as discussed in C5 of Chapter 2), sterile effects, and background conditions. These experimental strategies,

[24] I have not discussed the possibility that the boundaries of mechanisms might be defined by grouping together those items that were selected for the performance of some function. This answer, however, presupposes a solution to the problem under consideration here. To establish that a component was selected for its contribution to a function, one must first show that the part contributes to the function and so is relevant in the sense under discussion here. (As a purely epistemic point, neuroscientists often know very little about how features of the brain evolved.)

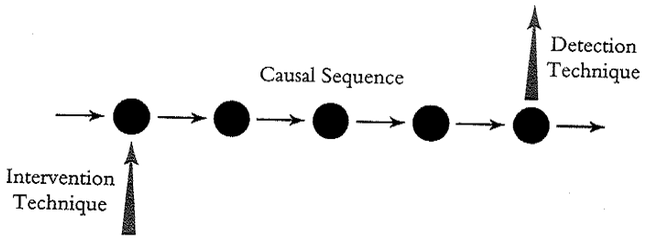

Figure 4.7a. Abstract representation of an experiment for testing etiological (causal) relevance

and their various well-known weaknesses, provide a valuable window on the norms of constitutive relevance.

To start with the more familiar case, etiological causal claims are tested with experiments of the sort diagrammed in Figure 4.7a. That figure represents an intervention (I) into a causal sequence to change variable (X) and a detection technique (D) that monitors the consequences (if any) of that intervention on some downstream variable (Y). In interlevel experiments, in contrast, the intervention and detection techniques are applied to different levels of mechanisms. (For present purposes, X's ϕ-ing is at a lower level than S's ψ-ing if X's ϕ-ing is a component of S's ψ-ing.) *Interlevel* experiments test the relationship between the components of a mechanism (the entities, activities, and organizational features at the lower level[25]) and the *explanandum phenomenon* (at the higher level). The left side of Figure 4.7b shows a bottom-up experiment, in which one intervenes to change a component in a mechanism (X's ϕ-ing) and detects changes in the behavior of the mechanism as a whole (S's ψ-ing). The right side of Figure 4.7b shows a top-down experiment, in which one intervenes to manipulate the phenomenon (S's ψ-ing) and detects changes in the activities or properties of the components in the mechanism (X's ϕ-ing).

Let me clarify the relationships involved in such experiments. X's ϕ-ing is a component in S's ψ-ing. S's ψ-ing can be understood as a complex input–output relationship. The inputs include all of the relevant conditions required for S to ψ. In the case of the action potential, this includes the

[25] By "level" in this context I mean the relationship between a mechanism as a whole and the entities, activities, properties, and organizational features of the mechanism taken individually. See Chapter 5.

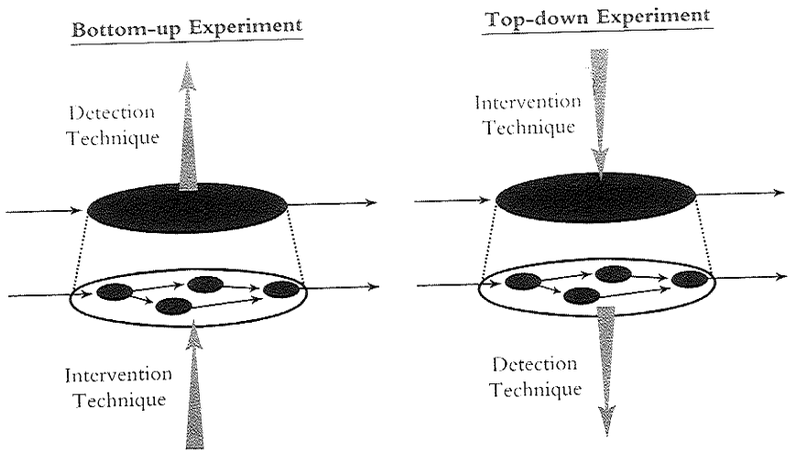

Figure 4.7b. Abstract representation of experiments for testing constitutive (or componential) relevance

stimulus delivered to the pre-synaptic cell and a host of other conditions of the sort described in the methods sections of scientific papers. The output is an action potential. Between these inputs and outputs is a mechanism, an organized collection of parts and activities. X is one of those parts, and ϕ is one of those activities. One intervenes on S's ψ-ing by intervening to provide the conditions under which S regularly ψs. Top-down experiments intervene in this way. Bottom-up experiments involve intervening into the components of the intermediate mechanism. Often they also involve putting S in the conditions for ψ-ing in order to see whether the intervention into the part changes whether S ψs or the way that S ψs. In each case, the goal is to show that X's ϕ-ing is causally between the inputs and outputs that constitute S's ψ-ing.[26]

Three varieties of interlevel experiment are common in contemporary neuroscience: interference experiments, stimulation experiments, and activation experiments.[27] They differ depending on whether the experiment

[26] I stress again that this relationship should not be understood causally. Nor should it be understood as a relationship between a supervenient event or property and its supervenience base. Rather I am talking about a relationship between a component in the mechanism and the behavior of a mechanism as a whole.

[27] There is a fourth kind of interlevel experiment, deprivation experiments, which I neglect here because they are so rare in neuroscience. In such experiments, one inhibits the behavior of a mechanism as a whole and detects changes in the behaviors of the parts. I am thinking, for example, of the experiment in which David Hubel sutured the eyes of kittens and monkeys to observe how the cortex develops when deprived of visual input.

is top-down or bottom-up, and on whether the intervention is excitatory or inhibitory. To reveal the criteria for assessing constitutive relevance, I examine some inferential challenges that these different kinds of experiment face. Neuroscientists are aware of these challenges. I mention them not as objections to interlevel experiments (which I take to be indispensable), but as data points in building a descriptively and normatively adequate account of constitutive relevance. A complete discussion of these methods, their strengths and weaknesses, and their relationships to one another is much needed but beyond the scope of this book. What follows is a skeletal framework that can be used for that purpose.

8.2.1. Interference experiments Interference experiments are bottom-up inhibitory experiments. They are represented on the left side of Figure 4.7b. In interference experiments, one intervenes to diminish, disable, or destroy some putative component in a lower-level mechanism and then detects the results of this intervention for the *explanandum phenomenon*. The assumption is that if X's ϕ-ing is a component in S's ψ-ing, then removing X or preventing it from ϕ-ing should have some effect on S's ability to ψ. In the simplest case, removing X or preventing it from ϕ-ing would eliminate or inhibit S's ψ-ing. If X is an inhibitory component, then intervening to remove X or to inhibit its ϕ-ing might produce or augment S's ψ-ing. The point of an interference experiment is to show that one can change S's ψ-ing by intervening to manipulate X's ϕ-ing.

Lesion experiments, for example, are interference experiments in which something intervenes to remove a portion of the brain and one then detects the effects of the lesion on task performance (see, for example, von Eckardt Klein 1977; Glymour 1994; Bub 1994). Clinical case studies, such as the cases of Leborgne (Broca 1861), H. M. (Scoville and Milner 1957), and Phineas Gage (Harlow 1868), are dramatic examples of interference studies. Interference experiments have also been crucial for discovering the mechanism of the action potential. When one intervenes to introduce mutations to the primary structure of the S_4 region, or to cleave the ball-and-chain inactivation gate in the Na^+ channel with proteolytic enzymes, or to inhibit channels with TTX or TEA, and then detects the effects of

the intervention on the shape of the action potential, one is conducting an interference experiment.

It is well known among scientists that interference experiments face significant challenges. One is that the mechanism sometimes compensates for the intervention. A second is that the intervention can sometimes influence the behavior of the mechanism as a whole indirectly.

(F1) Compensation: there are circumstances under which S's ψ-ing does not change after X and its ϕ-ing are disrupted, even though X and its ϕ-ing are relevant to S's ψ-ing. X could be redundant. The work of one kidney, or of one bilateral brain region, can sometimes be assumed by its partner with no diminution of function. In other cases, the mechanism might compensate for the loss of a part by recovering (healing the part), by making new use of other parts, or by reorganizing the remaining parts. Each of these possibilities is illustrated in people who have suffered strokes. Over the weeks following their stroke, many of the affected functions often return because the affected brain regions recover or reorganize, or because the person learns new ways to perform old tasks. The failure to see effects of interference is, as all neuroscientists know, insufficient to show that the part is irrelevant to the mechanism.[28]

(F2) Indirect interference: there are also circumstances in which interfering with X's ϕ-ing can change S's ψ-ing even though X's ϕ-ing is irrelevant to S's ψ-ing. For example, a brain lesion can disrupt the blood supply to surrounding brain regions, or it can produce swelling in the surrounding tissue that disrupts normal functioning in those areas. In these cases, the lesion delivered to brain region X has indirect effects on other areas, and those indirect effects are responsible for the observed deficit in S's ψ-ing. For example Anand and Brobeck (1951) report that lesions to the lateral hypothalamus stop rats from eating. They conclude that the lateral hypothalamus is a hunger center. Subsequent research confirms that the rats stop eating. They also stop *moving*. Electrolytic lesions to the lateral hypothalamus damage not only indigenous cells, but also a pathway of

[28] Compensatory responses are frequently incomplete. I am assuming the worst-case epistemic scenario for the neuroscientists and, conversely, the best-case recovery scenario for the patient, in order to make the interpretive challenge as stark as possible. Thanks to John Bickle for urging me to make this point.

neurons passing through the hypothalamus (the nigrostriatal bundle) that is thought to be a component in mechanisms regulating general arousal.[29] Again, neuroscientists are aware of this problem; Anand and Brobeck's paper would not be accepted for publication in any contemporary neuroscience journal because it does not conform to the norms that have subsequently evolved for evaluating interference experiments. In cases of this sort, however, one intervenes to change X and detects a change in S's ψ-ing, although the observed relationship is not due to the fact that X is a component, but rather to the fact that the disruption of X changes A, and A is a component in the mechanism of S's ψ-ing.[30]

An adequate account of constitutive relevance should help us to understand how each of these interpretive difficulties (compensation and indirect interference) is met. I return to this in the next major section.

8.2.2. Stimulation experiments Stimulation experiments are bottom-up, excitatory experiments. They are represented along with inhibition experiments on the left side of Figure 4.7b. In stimulation experiments, one intervenes to excite or intensify some component in a mechanism and then detects the effects of that intervention on the *explanandum phenomenon*. The assumption is that if X's ϕ-ing is a component in S's ψ-ing, then one should be able to change or produce S's ψ-ing by stimulating X. In the clearest case, one could make S ψ by making X ϕ. If X and its ϕ-ing play an inhibitory role in the mechanism, then stimulating X to ϕ would diminish or eliminate S's ψ-ing. If X or its ϕ-ing has only a modulatory role in S's ψ-ing, then stimulating X would change S's ψ-ing.

The classic example of stimulation experiments is Gustav Fritsch and Eduard Hitzig's (1870) work on the motor cortex (see Bechtel forthcoming). Fritsch and Hitzig performed a series of experiments on dogs in which they delivered low-grade electrical stimuli to a cortical area now known as the motor strip (see Bechtel forthcoming). Localized stimuli along

[29] One can dissociate these possibilities by using techniques that kill the dopaminergic fibers passing through the hypothalamus but that leave the indigenous cells intact and vice versa.

[30] Again, I am not arguing for skepticism about the results of these experiments. There are standards for distinguishing good interference experiments from bad, and one task of the philosophy of neuroscience is to make those explicit and to justify them. Here, I am using the well-known problems of interference experiments as a basis for showing what neuroscientists mean when they say that a part is a component in a mechanism. I am merely explaining why neuroscientists often claim that mere lesion experiments are insufficient for that purpose.

this area produce regular and repeatable movements in specific muscles, including the legs, the tail, and the facial muscles. The ability to produce focal movements predictably by stimulating areas of the brain is potent evidence that the stimulated area plays a role in motor mechanisms. Many of the electrophysiological experiments leading to the discovery of the mechanism of the action potential involve stimulating cells by injecting current.

Stimulation experiments give rise to interpretative complexities similar to those generated by interference experiments, as shown below (see Bechtel and Stufflebeam 2001 for examples).

(S1) Compensation: just as a neural mechanism can sometimes recover from interference, it can also sometimes recover from stimulation. Stimulating X to ϕ thus might not lead to S's ψ-ing even though X's ϕ-ing is relevant to S's ψ-ing. For example, homeostatic mechanisms might work to "siphon off" the stimulation or to adjust activities elsewhere in the mechanism to compensate for its effects. One example of such compensatory responses is drug tolerance, in which repeated exposure to a drug might lead to the need for larger doses to achieve the required effect. Tolerance to morphine is thought to result from compensatory responses within the endogenous opioid receptors, and the diminishing returns from L-Dopa in the treatment of Parkinson's disease are thought to result, at least in part, from downregulation of dopamine receptors in the basal ganglia. In most cases of stimulation, such compensatory responses are sufficiently delayed that they pose no threat to the interpretation of controlled experiments that test for a drug's effect, but it is not always possible to rule out short-term compensatory responses that would not be so evident.

(S2) Indirect effects: another challenge facing stimulation experiments arises from the possibility of indirect effects of the stimulation. For example, the stimulation delivered to X might spread to some other component B, where B is a component of S's ψ-ing. In that case, one can manipulate S's ψ-ing by manipulating X's ϕ-ing, but X's ϕ-ing is not a component in the mechanism for S's ψ-ing. Fritsch and Hitzig worried that their stimuli spread to other portions of the cortex. Subsequent experimenters refined the intensity of the electrical stimulus to localize the effects of the stimulation to just the brain regions under study. Similar refinement has taken place in experimental protocols involving pharmacological agents and genetic manipulations. One goal in designing a good stimulation experiment is to confine the stimulus to just the putative component or property under study.

8.2.3. Activation experiments The last kind of interlevel experiment is activation experiments. In activation experiments, one intervenes to activate, trigger, or augment the *explanandum phenomenon* and then detects the properties or activities of one or more putative components of its mechanism. These excitatory, top-down experiments are represented on the right side of Figure 4.7b. The basic assumption behind activation experiments is that if X is a component in S's ψ-ing, then there should be some difference in X depending on whether S is ψ-ing or not. In the most intuitive case, X would become active, or would increase its activity from baseline, when S begins to ψ. In parallel with cases of omission and prevention, however, it is also possible that X's ϕ-ing inhibits S's ψ-ing, and that activating S's ψ-ing therefore attenuates or eliminates X's ϕ-ing. Regardless, the point of an activation experiment is to show that interventions that change S's ψ-ing are accompanied by changes in X's ϕ-ing.

There are several common varieties of activation experiment at all levels in neuroscience. In PET and fMRI studies, one activates a cognitive system by engaging the experimental subject in some task while monitoring the brain for markers of activity, such as blood flow or changes in oxygenation. (For philosophical discussion of these techniques, see, for example, Bechtel and Stuffelbeam 2001; Bogen 2001, 2002. For a state of the art look at the techniques and its challenges, see Raichle and Mintun 2006). In single- and mutli-unit recording experiments, one engages the subject in a task while recording the electrical activity in neurons. In other studies, researchers monitor the production of proteins, or the activation of immediate early genes such as c-fos and c-jun. The experiments leading up to Hodgkin and Huxley's model of the action potential involved generating action potentials and monitoring single ionic currents while the neuron spiked. Activation experiments also face inferential perils, as described below.

(A1) Mere correlates: one challenge for activation experiments is that the activated component might be a mere correlate of the phenomenon. For example, engaging a subject in a cognitive task increases blood flow to brain regions activated by the task. PET researchers routinely take the increase

in blood flow as a marker of activity in components, but no researcher believes that the increase in blood flow is itself part of the mechanism for such cognitive tasks. Instead, the changes in blood flow are treated as poorly understood background conditions rather than as established components in the mechanism under study.[31] More generally, intervening to make S ψ might activate some component X of S, but the activation of that component has sterile effects, relative to S, on some irrelevant part, C. C would then be strongly correlated with task activation, but it would not be part of the mechanism. The lesson is that compelling top-down results, while an important part of establishing constitutive relevance, cannot alone establish constitutive relevance.

(A2) Tonic contributions: a major assumption of activation experiments is that X and its ϕ-ing must change during S's ψ-ing. Yet it is possible that a component plays a static role in the mechanism. Consider, for example, the contribution of the non-channel regions of the membrane, or perhaps Schwann cells, to the action potential. There can be no potential difference without a membrane that is largely impermeable to ions. Although channels change the permeability of the membrane, other portions of the membrane remain crucially impermeant. The existence of insulating Schwann cells that wrap the axon of a nerve cell allow the action potential to propagate quickly along its length. Schwann cells do not change during the propagation of action potentials; their insulating effect is a static, or tonic, contribution (or at least they are often described this way).

These experimental strategies—interference, stimulation, and activation—cannot be understood fully in isolation. They are typically used in conjunction because the strengths of one strategy compensate for the weaknesses of the others.[32]

8.3 *Constitutive relevance as mutual manipulability*

The close analogy between causal experiments and interlevel experiments suggests that the manipulability account of etiological relevance might provide a model for thinking about constitutive mechanistic relevance. My

[31] If one were to cut off blood flow for very long, the brain region would no longer function, but that is not the point. I am referring to the increase in blood flow subsequent to activation.

[32] Philosophers of neuroscience have said very little about the structure and limitations of these experimental strategies; see Bogen 2002; Hardcastle 2002; Uttal 2001.

working account of constitutive relevance is as follows: a component is relevant to the behavior of a mechanism as a whole when one can wiggle the behavior of the whole by wiggling the behavior of the component *and* one can wiggle the behavior of the component by wiggling the behavior as a whole. The two are related as part to whole and they are *mutually manipulable*. More formally: (i) x is part of S; (ii) in the conditions relevant to the request for explanation there is some change to X's ϕ-ing that changes S's ψ-ing; and (iii) in the conditions relevant to the request for explanation there is some change to S's ψ-ing that changes X's ϕ-ing. This simple formulation needs considerable refinement.[33]

There are significant differences between etiological and constitutive relevance. Because X is part of S (and ϕ is part of ψ), the relationship between them is only uncomfortably viewed as causal. Constitutive relevance is symmetrical in a way that etiological (that is, causal) relevance typically is not. In constitutive mechanistic relations, one can change the *explanandum phenomenon* by intervening to change a component (as illustrated by interference and stimulation experiments), or one can manipulate the component by intervening to change the *explanandum phenomenon* (as illustrated by activation experiments). Although there are *some* cases of cause and effect variables in which the manipulability relationships are bidirectional (as in cases of feedback), many, if not most, causal relationships are unidirectional. In contrast, all constitutive dependency relationships are bidirectional. This is the core reason why constitutive relevance should be understood in terms of *mutual* manipulability rather than in terms of the unidirectional variety introduced in Chapter 3. Second, in the constitutive relation, a token instance of the property ψ is, in part, constituted by an instance of the property ϕ; as such, the tokening of ϕ is not logically independent of the tokening of ψ. At least since Hume, many philosophers have held that causes and effects must be logically independent. If one endorses this restriction on causal relations, then one should balk at positing a causal relationship between constitutively related properties. Finally, because the constitution relationship is synchronic, ϕ's taking on a particular value is not

[33] This should not be confused with a claim about supervenience. Supervenience, in this case, amounts roughly to the claim that there can be no difference in S's ψ-ing without a difference in the mechanism for S's ψ-ing. Supervenience so stated is a relation between a phenomenon and the corporate behavior of the organized components. The relevance relation, in contrast, holds between the phenomenon and one of the components. The supervenience claim, note, is not symmetrical. I have no reason to deny weak and global forms supervenience, but that is not what I am discussing here.

temporally prior to ψ's taking on its value.[34] If one is committed to the idea that causes must precede their effects, then constitutive relationships are not causal relationships. These differences warrant caution in thinking of constitutive (interlevel) relations as causal. It seems appropriate to acknowledge these differences by marking the linguistic distinction between causation and componency, and so between etiological relevance and constitutive relevance.

I return now to the proposed sketch of constitutive relevance. According to that sketch, X's ϕ-ing is constitutively relevant to S's ψ-ing if the two are related as part to whole and the relata are mutually manipulable. There should be some ideal intervention on ϕ under which ψ changes, and there should be some ideal intervention on ψ under which ϕ changes.

With respect to the first of these conditionals, an *ideal* intervention I on ϕ with respect to ψ is a change in the value of ϕ that changes ψ, if at all, *only via* the change in ϕ. This implies that:

- ($I1_c$) the intervention I does not change ψ directly;
- ($I2_c$) I does not change the value of some other variable ϕ^* that changes the value of ψ except via the change introduced into ϕ;
- ($I3_c$) that I is not correlated with some other variable M that is causally independent of I and also a cause of ψ; and
- ($I4_c$) that I fixes the value of ϕ in such a way as to screen off the contribution of ϕ's other causes to the value of ϕ.[35]

Consider these briefly. Requirement ($I1_c$) is intended to rule out cases in which the putative intervention on X and its ϕ-ing directly fixes the value of ψ. (Remember, ψ does not supervene on ϕ. Rather, ϕ is part of the mechanism for ψ-ing.) If one were testing whether Na^+ channels are relevant to changes in membrane voltage, and one intervened to activate Na^+ channels by raising membrane voltage, the observed change in membrane voltage might be due to the intervention rather than to the activation of Na^+ channels. Requirement ($I2_c$) excludes those cases of indirect effects mentioned in F2 and S2 above. In those cases, an intervention has indirect effects (ϕ^*) that account for the observed changes to ψ. ($I3_c$) is required to rule out cases in which the intervention is

[34] For a detailed discussion, see Kim (2000) and Craver and Bechtel (forthcoming).

[35] The numbering here is intended to parallel that for etiological causal claims introduced in Chapter 3.

correlated with other determinants of the value of ψ. For example, the control organisms in lesion experiments typically undergo sham surgeries to ensure that the observed effects are not due to anesthesia or other correlated aspects of the surgical procedure rather than the lesion. It is not intended to rule out cases in which, for example, M is causally intermediate between I and X, as ruled out by ($I2_c$). Finally, ($I4_c$) is required to ensure that the intervention in fact changes the value of ϕ as intended.

Consider the first of the two conditionals that constitute the mutual manipulability account, that which asserts a conditional relationship between low-level interventions and high-level consequences. Putting it roughly:

> (CR1) When ϕ is set to the value ϕ_1 in an ideal intervention, then ψ takes on the value $f(\phi_1)$.

CR1 reflects the importance of bottom-up experiments, such as interference and stimulation experiments, for testing claims of constitutive relevance. Let ϕ be a variable representing the activity of a brain region, and let ϕ_1 represent the activity produced by ablating the region (that is, ϕ_1 = off). If X's ϕ-ing is necessary for S's ψ-ing, then S should no longer ψ (that is, $f(\phi_1)$ = off). If one removes the ball-and-chain inactivation gate from Na^+ channels, then the channel should not longer inactivate.

CR1 must be further restricted to conditions germane to a request for explanation. This is required to accommodate the fact that in neuroscience and elsewhere, one is not interested in whether just any change to ϕ could change ψ, or in whether changes to ϕ under just any conditions could change ψ, but rather in whether the changes can be observed in conditions that are explanatorily salient. What counts as an experimentally salient condition should be judged on a case-by-case basis, but the general idea is that if one is trying to understand the way a mechanism works in a healthy organism, and one is positing a constitutive explanatory relationship that holds only under extreme laboratory or pathological conditions, then one will not have identified a component of the mechanism in explanatorily relevant conditions. If one is interested in explaining the behavior of a mechanism under diseased or industrial conditions, then one will be interested in componency relations under those conditions. Although we are often interested in states of health or features that have been selected for, there is no reason to insist upon this restriction. There is no way to

know what constitutes the "appropriate" conditions without specifying the pragmatic context in which one is operating.

One can exclude sterile effects and other mere correlates by requiring that the experiment satisfy CR1. Sterile effects are properties or behaviors of a component that are irrelevant to the behavior of a mechanism as a whole. Performance of cognitive tasks, for example, is routinely correlated with hemodynamic changes, but this does not mean that the hemodynamic changes are part of the mechanism involved in task performance (as all MRI researchers know). Hemodynamic changes can be ruled out as components of the mechanism on the grounds that intervening to prevent the increase in blood flow during a task will not prevent one from performing the task. Of course, preventing blood flow to a region can quickly degrade task performance, and perhaps preventing the increase in blood flow would have long-term consequences as well. However, because hemodynamic changes typically *follow* the performance of a task, it is safe to assume that preventing those changes cannot alter task performance. Most generally, CR1 excludes correlations from constitutive explanations because intervening to change a mere correlate will not alter the phenomenon. Knowing that one can manipulate S's ψ-ing by manipulating X's ϕ-ing in various ways allows one to say how S's ψ-ing is different when X is removed, or when X's ϕ-ing is altered. In other words, a relationship that satisfies CR1 allows one to answer a range of what-if-things-had-been-different questions about how the mechanism will behave under a variety of interventions into its components. Mere correlations across levels do not allow one to answer such a range of questions.

Nonetheless, satisfying CR1 is neither necessary nor sufficient for X's ϕ-ing to be relevant to S's ψ-ing. Consider these in turn.

Satisfying CR1 is unnecessary because compensatory responses (such as recovery, redundancy, and reorganization) can prevent changes to S's ψ-ing (as noted in (P1) and (S1) above). One way that scientists solve problems of this sort is to design experiments that avoid or prevent the compensatory response. They try to show that the intervention on X's ϕ-ing induces changes in S's ψ-ing if one detects S's ψ-ing before S has had time to recover, or if the other redundant components are occluded or taxed, or if one prevents the system from reorganizing. More formally, there should be an ideal intervention on X's ϕ-ing that changes the value of S's ψ-ing under the conditions (CR1a) that the intervention, I, leaves all

of the other dependency relations in S's ψ-ing unchanged and (CR1b) that other interventions have removed the contributions of other redundant components. CR1a rules out cases of recovery and reorganization in the mechanism. CR1b rules out cases of redundancy. These two conditions are not merely ad hoc additions to the account. They correspond to the kinds of experiment that researchers do to overcome these inferential challenges. Although a system might reorganize in response to an intense stimulus (as in cases of drug tolerance), such effects are often delayed, allowing researchers to observe short-term changes before recovery or reorganization is complete. In some cases, researchers may be able to intervene to prevent the system from reorganizing. Problems of redundancy, likewise, can be met with experiments that inhibit the redundant mechanisms and, thereby, unmask the causal contribution of the part in question. Even if removing one kidney has little physiological effect, removing the second has dire physiological consequences. A final way that experimentalists deal with this kind of problem is by intervening in a way that does not prompt the system to compensate. As I show below, activation experiments can be used to detect correlated activity in multiple redundant parts, and they can usually be carried out in ways that do not prompt the system to reorganize. One of the virtues of such top-down experiments is that they sidestep inferential perils that bottom-up experiments cannot.

CR1 is also insufficient to establish a component's constitutive relevance because interventions into *background conditions* can change S's ψ-ing even though they are not part of the mechanism (see (F2) above). Lesioning the heart can prevent word-stem completion, but the heart is not part of the word-stem completion mechanism. In such cases, the lesioned or stimulated item *is* relevant to *explanandum phenomenon* in the sense that one can manipulate the phenomenon by intervening to change parts, but the parts are not components in the mechanism.

No doubt, the distinction between background conditions and components is often drawn on pragmatic grounds. However, such pragmatic decisions can be made on an objective base. Here are some ways of doing so. First, sometimes mere background conditions are identified by conjoining interference and stimulation strategies. Intervening to inhibit a background condition B's ϕ-ing may inhibit S's ψ-ing, but one cannot stimulate S's ψ-ing by stimulating B's ϕ-ing. For example, while interfering with the heart interferes with word-stem completion, one cannot produce

word-stem completion by stimulating the heart. Second, sometimes background conditions can be ruled out on the basis of activation experiments. Although one can interfere with S's ψ-ing by interfering with background condition B's ϕ-ing, at least in many cases, one cannot alter B's ϕ-ing by manipulating S's ψ-ing. For example, lesioning the heart might produce deficits in word-stem completion, but engaging a subject in word-stem completion will not change the behavior of the heart (except under torturous word-stem completion tasks outside of the context of the request for explanation). Third, the effects of interfering with background conditions tend to be nonspecific, that is, they affect many phenomena besides the one under study. Researchers learned, for example, that the lateral hypothalamus is not a hunger center by recognizing that the hypothalamic lesions prevent the animals from doing most of the things that animals do. Lesions to the heart would impair not only word-stem completion but also everything else distinctive of a living organism. Finally, the effects of interventions that change background conditions on the behaviors of mechanisms are often unsubtle. One cannot reliably produce subtle changes in word-stem completion by even arbitrarily subtle interventions to change the heart; interventions on the heart that have any effect seem to have switch-like effects. Slowing the heart, for example, will have no effect up to a threshold beyond which word-stem completion rapidly ceases. One who truly understood word-stem completion, however, if provided with the appropriate tools (a sizeable if), would be able to intervene into the mechanism to subtly manipulate the mechanism's output.[36] Criteria of this sort might provide a means for drawing the distinction between background conditions and components in a mechanism and for showing how CR1 might be supplemented to meet this problem case.

Part of the motivation for associating constitutive relevance with *mutual* manipulability is that bottom-up and top-down experiments are mutually reinforcing in the search for components in a mechanism. The inferential complexities involved in interpreting one such experimental strategy are often resolved by applying another strategy. None of the strategies is, by itself, sufficient to establish the constitutive relevance of a putative component. The strategies cannot be assessed fully in isolation from one another.

[36] James Woodward mentioned the third and fourth of these criteria in personal conversation.

For this reason, the mutual manipulability account contains the second conditional:

(CR2): if ψ is set to the value ψ_1 in an ideal intervention, then ϕ takes on the value $f(\psi_1)$.

CR2 is intended to describe the effects of top-down experiments, such as the use of functional imaging or the use of other biological markers of activity. One compares, for example, brain scans taken during a task (ψ_1) to brain scans during rest (ψ_2) in order to see which areas of the brain change across these two conditions. One compares, for example, cfos and cjun expression in cells during a task and not during the task to see if the cells are producing proteins relevant to the task. As discussed in the preceding section, these experimental strategies are common and useful.

Nonetheless these strategies face the challenges of mere correlation and tonic activation. Blood flow might increase during a task even if the increase in blood flow is not part of the mechanism performing the task. The tonic activity of a part might be relevant to performance of a task even if task performance does not change its level of activity (that is, not all parts of a mechanism have to change when the mechanism is working). In practice, these problems can be overcome by bottom-up experiments. Mere correlates of task performance cannot be manipulated to change task performance, but task performance can be manipulated by manipulating tonic activities. This is why I argue that constitutive relevance should be understood as *mutual* manipulability. What the second conditional adds, as I argued above, is a tool for dealing with problems of compensatory responses and for sorting components from background conditions. For these reasons, in addition to recognizing that such experiments are crucial in contemporary neuroscience, I include CR2 in the account of constitutive relevance.

In sum, I conjecture that to establish that X's ϕ-ing is relevant to S's ψ-ing it is sufficient that one be able to manipulate S's ψ-ing by intervening to change X's ϕ-ing (by stimulating or inhibiting) and that one be able to manipulate X's ϕ-ing by manipulating S's ψ-ing. To establish that a component is irrelevant, it is sufficient to show that one cannot manipulate S's ψ-ing by intervening to change X's ϕ-ing and that one cannot manipulate X's ϕ-ing by manipulating S's ψ-ing. The complexities in the componency relationship make it difficult to say more about the intermediate cases in which only one half of the mutual manipulability

account is satisfied. What to say in such cases, I suspect, depends on details peculiar to given experiments that admit of no general formulation. Nevertheless, the mutual manipulability approach is a suitable starting point for an account of constitutive relevance.

Relationships of mutual manipulability can and should replace the requirement of derivability as a regulative ideal on constitutive explanations in neuroscience. One need not be able to derive the phenomenon from a description of the mechanism. Rather, one needs to know how the phenomenon is situated within the causal structure of the world. That is, one needs to know how the phenomenon changes under a variety of interventions into the parts *and* how the parts change when one intervenes to change the phenomenon. When one possesses explanations of this sort, one is in a position to make predictions about how the system will behave under a variety of conditions. Furthermore when one possesses explanations of this sort, one knows how to intervene into the mechanism in order to produce regular changes in the phenomenon. Explanations in neuroscience are motivated fundamentally by the desire to bring the CNS under our control. The mutual manipulability account of constitutive relevance makes that connection explicit. Finally, the possibility of multiple realization does not even arise for the mechanistic account. It is not required that all instances of ψ-ing be explained by the same underlying mechanisms. What matters is that each instance of ψ-ing is explained by a set of components that are relevant to ψ in that particular mechanistic context. There are no doubt epistemic difficulties of determining when two mechanistic contexts are equal, but there is no conceptual difficulty seeing how the same type of phenomenon could be explained by different components in different contexts.

9. Conclusion

Both the systems tradition and the reduction tradition share a common goal of understanding constitutive explanation—that is, of understanding how the behavior of a whole is explained in terms of the behavior of its parts. For reasons that I discuss in the introduction to this chapter, most reductionists have now abandoned the classical model of reduction (sometimes called "strong reduction") according to which constitutive explanations involve forming transtheoretic identities and deriving one theory from another.

Those who have abandoned strong reduction commonly replace it with a weaker alternative, according to which reduction merely involves explaining higher-level phenomena in terms of underlying mechanisms (Kim 1989; Sarkar 1992; Smith 1992; Wimsatt 1976b). This move comes at a cost. When reductionists abandon strong reduction they also abandon the model of explanation that lies at its heart: namely, the CL model and the nomic expectability thesis. The problem is that there is no available account of constitutive mechanistic explanation to take its place.

My causal-mechanical account of constitutive explanation is a restricted and elaborated variant of accounts developed within the systems tradition (especially those found in the work of Bechtel, Cummins, Haugeland, Simon, and Wimsatt). The primary worry about previous such models is that they focus more on *describing* mechanistic explanations than they do on revealing the *norms* by which mechanistic explanations are and should be assessed. My friendly criticisms of Cummins's model are intended to illustrate how an accurate description of constitutive explanation can fall short of satisfying this normative objective. Because of their limitations as normative models, the models of the system tradition are not yet suitable competitors to classical reduction, the primary value of which is that it provides a regulative ideal on explanations. If the systems tradition is to challenge classical reduction as an account of constitutive explanation, it must provide an alternative regulative ideal. In this chapter, I take some steps towards rectifying that problem as well.

To see the progress that has been made, let us ask: How must Cummins-style functional analysis be restricted to provide a normatively adequate account of mechanistic explanation?

First, one needs to add the core normative requirement that mechanisms must account fully for the *explanandum phenomenon*. Ideally, it is not enough to account for just normal input–output conditions. One must also account for the multiple features of a phenomenon, including its precipitating conditions, manifestations, inhibitory conditions, modulating conditions, nonstandard conditions, and byproducts. Good explanations account for all of the features of a phenomenon rather than a subset.

Second, one needs to add that mechanistic explanations are constitutive. They explain the behavior of the mechanism as a whole in terms of the activities of its component parts. The parts should not be mere how-possibly fictions. Instead, they should exhibit clusters of properties, they

should be robustly detectable, they should be able to be used for purposes of intervention, and that they should be physiologically plausible.

Third, one needs to add that the activities appealed to in a compositional analysis should satisfy the criteria discussed in Chapter 3. This addition is required to rule out mere time-courses, effect-to-cause pairs, effects of common causes, and irrelevant causes.

Fourth, one needs to add a notion of organization. Organization is not merely a matter of being describable in terms of a box-and-arrow diagram or a program. Instead, it involves the active, spatial, and temporal organization of different components. This addition is required to distinguish mechanistic explanations from aggregate explanations, morphological explanations, and taxonomies.

Finally, one needs to supplement functional analysis with an account of constitutive relevance. Without such an account, functional analysis fails to offer an alternative to reduction, and it does not have the resources to exclude irrelevant components from the mechanism. The mutual manipulability account is a plausible condition of constitutive relevance because it fits well with experimental practice and because it is an extension of the view of etiological relevance advanced in Chapter 3.

My emphasis on constitutive mechanistic explanation (and its status as a competitor to classical reduction) should not lead one to forget that I am primarily interested in defending a multilevel view of explanation. Constitutive explanation is one important kind of explanation in neuroscience. But saying so does not commit me to the view that all explanations are constitutive. Nor does it commit me to the fundamentalist view that all explanations are achieved by looking to the lowest possible levels. In the next three chapters, I develop a view of levels (Chapter 5), I argue against fundamentalist claims that causal explanations can be given only at the lowest possible levels (Chapter 6), and I argue that the unity of neuroscience is constructed in the effort to build multilevel mechanistic explanations (Chapter 7).

5

A Field-Guide to Levels

Summary

Explanations in neuroscience typically span multiple levels. The term level, however, is multiply ambiguous. I develop a taxonomy of different kinds of levels, and I show why one must be careful to keep these different kinds distinct. Using an example from contemporary neuroscience—the multilevel mechanisms of spatial memory—I argue that "levels of mechanisms" captures the central explanatory sense in which explanations in neuroscience (and elsewhere in the special sciences) span multiple levels. The multilevel structure of neuroscientific explanations is a consequence of the mechanistic structure of neuroscientific explanations. I emphasize the importance of levels of mechanisms by showing how other common notions of levels (such as levels of science, levels of theories, levels of control, levels of entities, levels of aggregativity, and mereological levels) fail to describe the explanatory levels appearing in the explanation for spatial memory.

1. Introduction

The descriptive fact that explanations in neuroscience typically span multiple levels gives rise to scientific disputes about the relative significance of different levels and to philosophical disputes about the existence and explanatory relevance of nonfundamental levels. Yet the term "level" is multiply ambiguous. Its application requires only a set of items and a way of ordering them as higher or lower. Not surprisingly, then, the term "level" has several common uses in contemporary neuroscience.[1] To

[1] Machamer (personal communication), Churchland and Sejnowski (1992), and Hardcastle (1998) called my attention to this fact.