Constitutive Explanatory Relevance[1]

Carl Craver

ccraver@artsci.wustl.edu

Washington University, St. Louis

1. Introduction. The problem of constitutive relevance is the problem of saying in what sense the components of a mechanism are explanatorily relevant to the behavior of a mechanism as a whole. My primary goal is to articulate this problem and to show that it must be solved if we are to understand mechanisms and mechanistic explanation. My secondary goals are to argue against some putative solutions to the problem (Section 2) and to sketch a positive account (Section 3). I argue that constitutive relevance can be clarified by investigating the relationships of mutual manipulability between the behavior of a mechanism as a whole and the properties and activities of its components. My approach is causal-mechanical in that it is a particular expression of Salmon's idea that explanations show how an explanandum phenomenon is situated within the causal structure of the world.

2. The Problem. The problem of constitutive relevance is related to a classic problem in philosophical discussions of the nature of explanation. According to the once-received covering-law model of explanation (Hempel 1965), to explain a phenomenon (the explanandum) is to show that its description follows from generalizations describing nonaccidental laws of nature coupled with descriptions of the relevant antecedent and background conditions (the explanans). The covering-law model is an epistemic conception of explanation, according to which explanations are arguments showing that one would be rational in expecting the explanandum phenomenon on the basis of one's knowledge of the laws of nature. Wesley Salmon (1984; 1989) and others argue convincingly that this epistemic model has difficulty satisfying scientific and common sense judgments about explanatory relevance. Here are some familiar examples:

> *The falling mercury on the barometer is explanatorily irrelevant to the impending storm despite the fact that storms regularly appear when barometric readings fall. (*Common Cause*).

[1] Portions of this talk were delivered to the Department of Philosophy at Lund University, the Department of Philosophy at the University of Cincinnati, and the Institut Jean Nicod. I would like to thank Max Kistler, Tony Landreth, Johannes Persson, and Petri Ylikoski for comments on earlier drafts.

*Jones' taking birth control pills is explanatorily irrelevant to his failure to get pregnant despite the fact that no males who take birth control pills get pregnant (*Irrelevancies*).

*The height of the flagpole and the elevation of the sun explain the length of the shadow, but the length of the shadow and the height of the flagpole do not explain the elevation of the sun (*Asymmetry*).

In each case, the explanandum phenomenon is subsumed under a general regularity. In each case, one would rationally expect the explanandum phenomenon on the basis of the cited explanans. Yet each case fails as an explanation either because it leaves out explanatorily relevant details (such as the behavior of barometers or the length of the shadow) or because it includes explanatorily irrelevant details (Jones' birth control pills).

Salmon uses these much-discussed examples to motivate a causal-mechanical account of explanation. Whereas covering-law explanations explain by showing that the explanandum phenomenon was to be expected on the basis of the laws of nature, causal-mechanical explanations explain by showing how something is situated within the causal nexus. This is an ontic view of explanation: explanation, and so explanatory relevance, is an objective feature of the causal structure of the world. It is not a feature of our psychology or of our inferential practices. Salmon views the causal structure of the world as a nexus of causal processes intersecting and exchanging marks (1984) or conserved quantities (1994). However, the ontic, causal-mechanical view of explanation need not be committed to so narrow a view of causation.

Salmon discusses two varieties of causal-mechanical explanation, corresponding to two ways of situating an item in the causal nexus. *Etiological* explanations explain a phenomenon by describing its antecedent causes. Such explanations are represented in the bottom half of Figure 1 (redrawn from Salmon 1984) as a set of causal processes terminating in E. The relevant features of E's past are (on Salmon's view of causation) literally connected to E via lines of physical transmission. Irrelevant features are not so connected. The falling mercury is not relevant to the weather because there is no causal process connecting one to the other. Jones' birth control pills do not intersect with any system that would, left to its own devices, produce a fetus. And the lines of production and propagation of causal influence run from the sun to the ground and not the other way round.

[Figure 1 Near Here]

Constitutive explanations are shown in the middle of Figure 1. E is again the explanandum, but in this case constituent, rather than antecedent, causal processes do the explaining. I follow Salmon for the moment in using the term constituent merely to describe the parts of something. In mechanistic explanations, the sort of constitutive explanation with which I am primarily concerned, the parts are components. Components are relevant parts. The notion of "part" can be taken in a very relaxed way to include any subvolume of the material in an object or any temporal slice of an activity or process. Components, in what follows, are a restricted set of the parts thus broadly construed, namely, those parts that are relevant to the explanandum phenomenon.

As an example, consider the mechanism for long-term potentiation (or LTP). Long-Term Potentiation is an increase in the efficacy of a synapse following rapid and repeated stimulation. It is explained by molecular mechanisms. The explanation of Long-Term Potentiation (or LTP) includes the activities of subcellular objects, such as calcium and magnesium ions, and NMDA receptors (for N-methyl D-aspartate, a pharmacological agent that acts on the receptor). Constitutive explanations are inherently interlevel explanations. The behavior of a mechanism as a whole is explained by the behaviors and properties of its components. The mechanism that regulates the opening and closing of the NMDA receptor, for example, is a component in the mechanism for LTP, and in that sense, it is at a lower level than LTP. Some parts of the cell are not relevant. Several ubiquitous intracellular enzymes, ions, and energy sources are not included in the explanation. There are no mitochondria. The apoptosis mechanism is not included. The problem of constitutive relevance is the problem of saying which parts are components in a mechanism and which are not.

Constitutive explanation is the primary focus of Chicago Mechanists such as Bechtel (Bechtel and Richardson 1993; Bechtel 2006); Darden (2006); Glennan (1996; 2002); Kauffman (1971); Lycan (1987); Machamer (Machamer et al. 2000), Richardson (Bechtel and Richardson 1993), Sarkar (1992), Wimsatt (1976). Although there are differences among us, the consensus view is that mechanisms are entities and activities organized such that they exhibit the phenomenon to be explained. The entities are the parts. The activities are what they do. Activities are the causal components of mechanisms. They are also temporal parts of mechanisms. Mechanistic explanations include the phenomenon to be explained and a description

of all of the parts, properties, activities, and organizational features that together exhibit the explanandum phenomenon, the behavior of the mechanism as a whole.

Salmon did not say what makes a constituent relevant to the behavior of a mechanism as a whole. His account of *etiological* relevance appeals to physical interactions between causal processes, and so it does not apply to constitutive explanatory relevance. The behavior of the mechanism as a whole is not a distinct causal process from the behaviors of the components. The space-time path of the mechanism as a whole includes the space-time path of its components. They coexist with one another, and so there is no possibility of their *coming to* spatiotemporally intersect with one another. If a conserved quantity is possessed by one of the parts (say, a certain mass or a charge), that conserved quantity is also possessed by the whole. If one of the parts bears a mark, that mark is always already born by the whole (by virtue of being born by one of its parts). Even if Salmon's transmission account can recover scientific and common sense intuitions about etiological relevance (for arguments that it does not, see Hitchcock 1995), it is a non-starter as an account of constitutive relevance.

Similar considerations can be generalized to argue that constitutive relevance should not be understood as a causal relationship at all. Many widely accepted views of causation are violated in the case of constitutive relationships. For example, no account of causation that requires causes and effects to be distinct and non-overlapping (such as Lewis 1983 and Kim 2000) can be applied without amendment to constitutive explanations, given that the behaviors of mechanisms as a whole and the behaviors of the parts overlap. If a view of causation is to cover cases of constitutive relevance, it will also have to abandon the common assumption that causes precede their effects, because the behavior of components occurs during the behavior of the mechanism as a whole. If a view of causation entails that causal relevance is asymmetrical, then it will be inapplicable to the relationship between components and the behavior of a mechanism as a whole. Causes produce their effects, and (at least in many cases) *not* vice versa. The relationship between a component and the behavior of a mechanism as a whole is always symetrical. Finally, if one views causation as a relationship of manipulability, then modularity constraints and restrictions on experimental interventions might be thought to prevent one from thinking of the interlevel relationship causally. One cannot, for example, intervene to change the value of the whole without thereby

changing the value of the components. (Later, I show how this can be understood without violating this constraint on interventions.)

The Chicago school has thus far said very little about explanatory relevance and nothing about constitutive relevance. Stuart Glennan (2002) argues, picking up on a suggestion from Woodward (2002), that the etiological relevance of one property or component of a mechanism to another can be understood usefully (although not reductively) in terms of the ability to manipulate one component by intervening on another. One cannot control storms by breaking barometers. One cannot increase Jones' chances of getting pregnant by replacing his birth control pills with placebos. And (at least in many cases) one cannot change the past by intervening in the present. This manipulationist view is promising in part because it so readily diagnoses and dispenses with these classic problems. That is progress. But Glennan, Woodward and the rest of the mechanists thus far continue Salmon's tradition of neglect for constitutive relevance. I will ultimately argue for a manipulationist view of interlevel relevance.

This problem is not unique to causal-mechanical views of explanation. Examples such as the barometer and the storm and Jones' birth-control pills have analogues that raise problems for constitutive covering-law explanations. Classical reduction, in which a revised and restricted reduced theory is derived from a reducing theory with the aid of bridge laws connecting the kind-terms in the two theories, is the covering law's view of constitutive explanation (see Schaffner 1993). There are spurious derivations that any adequate account must rule out.

First, the behavior of a mechanism as a whole might be merely correlated with some activity or behavior of a part. This can happen when the activity or behavior of a part is caused or otherwise modulated by activities in the mechanism (analogously to the barometer and the storm). For example, functional imaging techniques work because changes in blood flow regularly follow changes in neural activity. Changes in blood flow are, so far as we know, not part of the mechanism. Rather, they are correlates that can be used as indicators. One could in some cases derive facts about our cognitive activities from facts about patterns in cerebral blood flow, but this would not count as an explanation. The blood flow changes are irrelevant. They are sterile effects.

Second, covering law explanations are arguments, and irrelevant premises do not affect the strength of an argument. But, as Salmon notes, irrelevant information is deadly for explanations. (This is

analogous to the point made by Jones' birth-control pills). Assuming that one could derive that a cell would induce LTP from statements about its molecular constituents (an assumption that receives little support from contemporary neuroscience), one could also achieve the same derivation with an additional set of statements about anything you like. Newton's laws and the Darwinian theory of evolution could be conjoined to statements about molecular constituents of cells without affecting the ability to derive features of the LTP phenomenon. Hempel and Oppenheim (in Hempel1965) raise a related problem in their famous footnote 33.

3. Relevance and the Boundaries of Mechanisms. Why do we need an account of constitutive relevance? First, an account of constitutive explanation that does not include an account of explanatory relevance cannot distinguish good explanations (that is, those that contain all and only the relevant explanatory facts) from bad explanations (that is, those that leave out relevant facts or include irrelevant facts). If the philosophy of science has a normative role to play in discussions of explanation in sciences such as biology and neuroscience, then it cannot sidestep the problem of constitutive relevance.

Second, without a view of constitutive relevance, the mechanistic approach to explanation threatens to collapse back into the covering-law model. For what can it mean for the organized activities of components to exhibit (or account for) the behavior of the mechanism as a whole if not that one can derive (or otherwise infer) a description of the latter from a description of the former? A view of constitutive relevance can, I believe, elucidate the sense in which a mechanism *accounts for* its phenomenon without abandoning the general commitment to the causal-mechanical view of explanation.

Third, the very idea of a mechanism presupposes the idea of constitutive relevance. To see why, consider one important difference between mechanisms and machines. Machines typically contain parts and have properties that are not in their mechanisms. The hubcaps, the mud-flaps, the windshield, and the fuzzy dice are all parts of a fine automobile, but they are not parts of the mechanism that makes it run. They are not *relevant* to the running behavior of the car, It would therefore be a mistake to include them in the mechanism for the car's running. This difference between mechanisms and machines turns at least in part on the fact that all of the constituents of a mechanism are relevant to what it does (they are components), while only some of the parts of machines are relevant to what they do (namely, those that are components in one or more of its mechanisms).

This point is closely related to one made by Stuart Kauffman (1971) in a core article for Chicago mechanism. Kauffman argues that there are many ways to break a machine or organism into parts depending upon one's perspective on the system. First, one accentuates or neglects different constituents of the machine or organism depending on which explanandum phenomenon one sets out to explain. To explain the blood's circulation, we focus on valves, arteries, and hearts, and we neglect adipose tissue, intestines and toenails. To explain respiration, we focus on lungs and diaphragms, but we neglect the venous valves. Second, in focusing on different explanandum phenomena one will cut the organism (or system) into entirely different parts. Researchers focusing on the regulation of body fluid homeostasis, for example, divide organisms into two fluid compartments: intracellular and extracellular. The very task of decomposition, in other words, is always relative to a perspective, and when the goal is explanation, what counts as an adequate decomposition depends upon what one is trying to explain.

Kauffman's point and the disanalogy between mechanisms and machines help to show that what counts as a single mechanism— what is within the boundaries of the mechanism and what is outside— is determined by whether the part is relevant to the behavior of the mechanism as a whole. How might one define the boundaries of mechanisms without appeal to a notion of explanatory relevance?

One might associate the boundaries of mechanisms with *compartmental boundaries.* Some mechanisms are entirely compartmentalized within barriers such as the cell-nucleus, the cell membrane, and the skin. Transcription typically happens inside the nucleus, and translation occurs in the cytoplasm. But examples of this sort are misleading in two respects.

First, as an empirical matter, mechanisms frequently transgress compartmental boundaries, and they often must do so if they are to work properly. The mechanism of chemical transmission, for example, crosses at least two compartmental boundaries. It includes components located in the presynaptic cell, in the synaptic cleft, and in the post-synaptic cell. Some of the entities (such as the receptors and ion channels) span the boundaries between those compartments. One might say, informatively, that chemical transmission happens in the synapse, but synapses are functionally defined, and include all of those components relevant to the communication between cells. The same situation holds for the many cognitive mechanism that include entities, properties, and activities both inside and outside the brain. Mechanisms are not always neatly contained.

Second, the fact that mechanisms such as the mechanism of protein synthesis *are* contained does not show that the container defines (rather than merely correlates with) the boundaries of the mechanism. One can take the protein synthesis mechanism out of the nucleus and allow it to work in a cell-free medium. This fact, which was crucial to the discovery of the mechanism of protein synthesis, shows that the compartment can be removed from many mechanisms without altering the mechanism's behavior at all: all of the *relevant* parts are still in place. Of course, it is trivially true that all material mechanisms can be circumscribed within a spatial boundary or contained within a physical boundary, but the problem is to state a rule (or short of that, some normative guidelines) for deciding where that physical or spatial boundary should be drawn if it is to include all and only the components in a mechanism.

Cartesian mechanists also worried about how to define the boundaries of mechanisms (see Des Chenne 2001). Descartes at times favors principles of spatial organization: the parts of machines move together; they can be transported together from one place to another while maintaining fixed relative positions. Some late Cartesians require that the parts of the machine must be in physical contact with one another. Few contemporary scientists or philosophers require that the parts of mechanisms must be physically connected. The parts of the NMDA receptor work by double-prevention. Depolarization in the post-synaptic cell releases the magnesium blockade of the channel and allows calcium to diffuse into the cell. Such causation by double-prevention cannot be understood in terms of physical connection. This fact suggests that some relevant parts of a mechanism might not be physically connected with one another. Parts of mechanisms also often move in separate directions. Magnesium moves out of the cell. Calcium moves in. Glutamate departs the presynaptic cell and diffuses to the post-synaptic cell. Nitric oxide moves in the opposite direction. Third, mechanisms are sometimes spatially ephemeral; they work only as components happen to come into the appropriate spatial arrangement. Diagrams of LTP mechanism gloss over the fact that the intracellular fluid is a witches brew of molecules and cascades, and the relevant reactions for a given mechanism or cascade could happen anywhere in the cytoplasm. Such mechanisms lack stable spatial relations. They could not be picked up and carried from one place to the next. Spatial relations are insufficient to define the boundaries of mechanisms.

Finally, one might delimit the boundaries of a mechanism by appeal to the causal interactions among the components in a mechanism. Herbert Simon (1969), for example, argues that variables are

clustered into systems (which can be understood as mechanisms for present purposes) when their interactions with one another are stronger than are their interactions with variables outside of that set. Wimsatt (1976) holds that the boundaries of a mechanism are defined by the relative strengths of intra- and extra-systemic causal interactions plus a pragmatic factor for tolerance of error. Haugeland (1998) describes mechanisms as "relatively independent and self-contained composites of components interacting at interfaces" (215). By "relatively independent and self-contained" he means that they interact more often and more intimately with items inside the interfaces than with those outside (215). Grush suggests a plug-and-play criterion, according to which components are those that can be taken out of the mechanism and replaced with functional equivalents.[2]

There are two issues that complicate the idea that the boundaries of mechanisms are delimited by the relative strengths of intra- and extra- systemic interactions.

First, there are *background conditions*.[3] The beating of my heart and my ability to lay down memories 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 lay down new memories would rapidly deteriorate. The LTP mechanism requires sources of energy, the stability of the cell membrane, the temperature and pH of the intra-and extra-cellular fluids. Background conditions are commonly judged to be outside of the mechanism. Ideally, an account of the boundaries of mechanisms would make sense of this judgment.

Second, the criterion of strength of interaction faces the challenge of ruling out *sterile effects*. The influx of calcium into the post-synaptic cell changes the calcium concentration in the witches brew, thereby affecting every intracellular signaling mechanism that shares calcium levels as a rate-limiting step (a fact that makes the coordinated working of biochemical mechanisms seem miraculous at times). NMDA receptors exert a host of 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 like changes in cerebral blood flow, these activities are sterile in the mechanism: they either produce no changes in the other components of the

[2] Grush's (2003) view is interestingly different from the others, and I do not consider his rich suggestion in detail in this paper. Note, however, that what counts as "plug and play" depends in part on our skills as pluggers and players, and so on features of our cognitive systems rather than objective features of the world, as demanded by the ontic conception.

[3] Note that these cannot be defeated by treating the termination condition of the mechanism

mechanism, or the changes they do produce make no difference to LTP. In contrast, the opening of the NMDA receptor, and the movement of calcium ions are all tightly coupled in a way that does make a difference to action potentials.

To conclude, in my view, all of these approaches to defining the boundaries of mechanisms are looking for the wrong sort of criterion. The spatial and interactive boundaries of mechanisms are epistemically and ontically secondary to the delineation of *relevance boundaries*. The spatial boundaries sought by late Cartesians are those that circumscribe all and only the *relevant* entities and activities. The causal boundaries sought by Simon, Wimsatt, and Haugeland are those that circumscribe all and only the relevant causal interactions.

In fact, it is surprising that anyone could have thought otherwise. This conclusion is apparent even in the consensus definition of mechanisms: the mechanism includes the entities and activities that are organized *such that they exhibit the phenomenon*. There are no mechanisms simpliciter. There are only mechanisms *of* behaviors. The boundaries of mechanisms are drawn around all and only the entities, activities, and organizational features relevant to the behavior of the mechanism as a whole. This, it seems to me, is a step in the direction of a causal-mechanical alternative to derivability as a regulative ideal for explanation. *The regulative ideal is that the mechanism should describe all and only the components, activities, and organizational features that are consitutively relevant to the explanandum phenomenon.*

4. Constitutive Relevance. So what is constitutive relevance? I argued above that it cannot be understood in terms of contact action and transmission of marks or conserved quantities. And similar arguments show that no causal relationship is suitable for understanding interlevel relations. My working hypothesis, which is no doubt preliminary and in need of further refinement, is that that a component is relevant to a phenomenon when the component and the phenomenon are mutually manipulable.

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 instances of those strategies (Bechtel 2006; Bechtel and Richardson 1993; Craver 2001). These experimental strategies, and their various well-known weaknesses, provide a valuable window on the norms governing claims about constitutive relevance and so on the relations that such claims are about.

In experiments used to test claims about etiological relevance, one intervenes to change the value of a putative cause varable and one detects changes (if any) in the effect variable. Claims about constitutive relevance are tested with interlevel experiments. 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 constituent parts of a mechanism (the entities, activities, and organizational features at the lower level[4]) and the explanandum phenomenon (at the higher level). The left hand side of Figure 2 shows a bottom-up experiment, in which one intervenes into 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 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).

[Figure 2 Near here]

This way of speaking about interlevel experiments is intended to be understood as follows: 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 LTP, this will include the stimulus delivered to the presynaptic cell and other conditions of the sort described in methods sections of scientific papers. The output is a potentiated synapse. 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 sufficient for S to ψ. Top-down experiments intervene in this way. Bottom-up experiments involve intervening into the components of the intermediate mechanism. 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.

There are three common varieties of interlevel experiment: interference experiments, stimulation experiments, and activation experiments.[5] They differ depending on whether the experiment is top-down or

[4] 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 Craver (forthcoming, chapter 5).

[5] There is a fourth kind of interlevel experiment, deprivation experiments, that I neglect here because they are so rare in neuroscience. In such experiments, one inhibits the behavior of a mechanism as a whole and

bottom-up, and on whether the intervention is excitatory or inhibitory (See Bechtel and Richardson 1993). Start with the bottom-up experiments.

Interference Experiments. Interference experiments are bottom-up inhibitory experiments. 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. Examples of interference experiments include lesion studies, deficit studies, gene knockouts, and the use of pharmacological antagonists. One might, for example, remove the NMDA receptor from the mouse and then see if the rat can learn to run a maze.

Stimulation Experiments. Stimulation experiments are bottom-up, excitatory experiments. 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 thought 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. Examples of stimulation experiments include brain stimulation studies, transgenic modification, and the use of pharmacological agonists.

It is well known among scientists that interference and stimulation experiments face two 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.

Compensation. There are circumstances under which S's ψ-ing will 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 compensates for the loss of a part by recovering, by making new use of other parts, or by reorganizing the remaining parts. The failure to see effects of interference is, as all scientists know, insufficient to show that the part is irrelevant to the mechanism.

---

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.

Mechanisms also sometimes compensate for stimulation. 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 in endogenous opioid receptors. 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.

Indirect Effects. The second challenge faced in such bottom-up experiments is to determine exactly how the intervention changes S's ψ-ing. There are 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, Anand and Brobeck (1953) 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 in fact do stop eating. They also stop *moving*. The lesions to the lateral hypothalamus, it turns out, damage not only indigenous cells, but also a pathway of neurons passing through the hypothalamus (the nigrostriatal bundle) that is thought to be a component in mechanisms regulating general arousal. In cases of this sort, however, one intervenes to change some putative component A and detects a change in S's ψ-ing, but the observed relationship is not due to the fact that A is a component, but rather to the fact that the disruption of A changes X, and X is a component in the mechanism of S's ψ-ing.

An analogous problem confronts stimulation experiments. For example, the stimulation delivered to A might spread to X. In that case, one can manipulate S's ψ-ing by manipulating A, but A 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. One goal in designing a good stimulation experiment is to confine the stimulus to just the putative component under study.

Activation Experiments. The last kind of interlevel experiment is the activation experiment. 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 this figure. 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. The point of activation experiments is to show that interventions that change S's ψ-ing are accompanied by changes in X's ϕ-ing. Examples of activation experiments include fMRI studies and the use of host of other markers for the activity of components in a mechanism.

Neural imaging studies have received sustained attention lately by methodologists in the cognitive neurosciences, and there is a great deal to be said about how these experiments work and how they fail. Here, I want to focus on two potential shortcomings.

Mere Correlates and Sterile Effects. The most obvious worry about many activation experiments is that the activated component may 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 I know of no researcher who believes that the increase in blood flow is itself part of the mechanism for such cognitive task. Instead, the changes in blood flow are treated as poorly understood background conditions rather than as established components in the mechanism under study.[6] As noted above, this is an example of a sterile effect. The lesson is that compelling top-down results, while an important part of establishing constitutive relevance, cannot alone establish the constitutive relevance of a component.

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 ATP production, to Long-Term Potentiation. 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. They do not change during the propagation of action potentials, or at least they do not change in a way that contributes to the mechanism. Their contribution is tonic.

[6] 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. For a lucid discussion of this indicator, see Raichle and Mintun (2006).

The crucial point is that these three experimental strategies—interference, stimulation, and activation— cannot fully be understood in isolation. These strategies are typically used together because the strengths of one strategy often compensate for the weaknesses of another.

This fact is built into my causal-mechanical approach to constitutive relevance. The close analogy between causal experiments and interlevel experiments suggests that the manipulability account of etiological relevance (developed along the lines of Woodward's 2003 account) might be extended to provide an account constitutive mechanistic relevance. My working account of constitutive relevance is as follows: a component is relevant to the behavior of a mechanism as a whole when the two are mutually manipulable— that is, 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. Somewhat more precisely, in the conditions relevant to the request for explanation, (i) there is some change to X's ϕ-ing that changes S's ψ-ing, and (ii) there is some change to S's ψ-ing that changes X's ϕ-ing. The relationship is symmetrical to reflect the fact that bottom-up and top-down experiments are mutually reinforcing in that they correct for inherent limitations in the others. (Note that this is not a supervenience or identity claim. It is a relationship between a whole and one of its parts. S's ψ-ing does not supervene on X's ϕing. Rather, it supervenes on the organized activities of all of the components in the mechanism.)

To begin with the bottom-up case, interference and stimulation experiments help to resolve the two worries just mentioned about activation experiments.

First, interference and stimulation studies allow one to distinguish mere correlates and sterile effects from relevant components. One cannot change the behavior of the mechanism as a whole by intervening to excite or inhibit lower-level correlates, but one can change the behavior of the mechanism as a whole by intervening to manipulate lower-level components. Performance of cognitive tasks, for example, is 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, such experiments exclude 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 (i) allows one to answer a range of what-if-things-are-different questions about how the mechanism will behave under a variety of conditions.

Second, while top-down activation experiments are blind to tonically active parts, removing or interfering with those parts nonetheless change the behavior of the mechanism as a whole. Breaking down the bilipid membrane of a cell will quickly reveal the contribution that it makes to the cell's electrical activities.

Activation experiments also help to resolve the problems associated with bottom-up experiments, namely the problems of compensation and indirect effects. Both interference and stimulation experiments suffer from the worry that the behavior of the mechanism during the bottom-up intervention is different from the way that the intact and unmolested mechanism behaves. One can avoid the problem of contemporary responses by doing nothing that would damage the system or cause it to repair itself. One activates the system under its standard conditions, and one watches how the mechanism behaves intact. Furthermore, one can check to see if stimulation experiments and lesion experiments have disturbed the mechanism through indirect effects by seeing whether the putative component is active when the mechanism is working in its normal conditions. Many of the critics of neuroimaging techniques, with whom I share a number of sympathies, forget that before neuroimaging was invented, there were always lingering doubts about how to interpret lesion and stimulation studies. One was always left wondering if the changes produced by intervening into the mechanism produced those changes in the same way that the mechanism works under standard, non-invasive, conditions. Imaging techniques allow one to watch the brain in action without such potentially destructive interference, and so help to alleviate some of these interpretive worries.

The main point is that none of these experimental strategies is alone sufficient to establish that a component is constitutively relevant to an effect. However, taken together, these experimental strategies are

powerful in the search for relevant components. The mutual manipulability account is designed to make this relationship more explicit.

I want to be clear about the limits of my conjecture. My claim is 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, and I argued that it captures many of the norms implicit in the practice of assessing interlevel relationships in the sciences.

Let us now return to the issue of background conditions and see how a manipulationist approach might be used to locate them outside the boundaries of mechanisms. Part of the problem, of course, is that background conditions are relevant to the behavior of the mechanism as a whole. Nonetheless, there are objective considerations that allow one to locate them as in some other sense outside of the boundaries of the mechanism proper.

Sometimes mere background conditions are identified by conjoining interference and stimulation strategies. Intervening to inhibit a background condition (B's) ϕ-ing might inhibit S's ψ-ing, but one cannot stimulate S's ψ-ing by stimulating B's ϕ-ing. For example, although 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 range of conditions relevant in the implied context of the request for explanation).

Third, the effects of interfering with background conditions tend to be non-specific, 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 plummets. 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 manipulate subtly the mechanism's output.[7] Criteria of this sort provide an objective basis on which to distinguish background conditions from components in a mechanism.

5. Conclusion. To conclude, I argue that neither the causal-mechanical view nor the once-received covering law account of explanation currently has an adequate approach to constitutive explanatory relevance. Such an account is required to distinguish good mechanistic and reductive explanations from bad because good explanations are those that include all and only the relevant components. Such an account is also required to delineate the boundaries of mechanisms, and so to say what a mechanism is. My proposal is that relationships of manipulability should replace the requirement of derivability as a regulative ideal for 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 would change under a variety of interventions into its 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. More importantly, however, when one possesses explanations of this sort, one knows how to intervene into the mechanism in order to produce regular changes in the phenomenon (and vice versa). The fine details of this view of explanation remain to be

[7] James Woodward mentioned the third and fourth of these criteria in personal conversation. He also convinced me to give up the search for necessary and sufficient conditions for ruling out background conditions as components in a mechanism.

worked out, but perhaps I have said enough to show that the mutual manipulability account has potential as a thoroughly causal-mechanical account of constitutive relevance.

Bibliography

Bechtel, W. (2006). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology.* Cambridge: Cambridge University Press.

——— Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Craver, C.F. (2001). 'Interlevel Experiments and Multilevel Mechanisms in the Neuroscience Of Memory'. *Philosophy of Science* (Supplement), 69: S83-97.

——— (Forthcoming). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

——— Bechtel, W. (forthcoming). 'Top-down Causation without Top-down Causes'. *Biology and Philosophy*.

Darden, L. (2006) *Reasoning in Biological Discoveries.* Cambridge University Press.

Des Chene, D. (2001). *Spirits & Clocks: Machine & Organism in Descartes*. Ithaca, NY: Cornell University Press.

Glennan, S. S. (1996). 'Mechanisms and the Nature of Causation'. *Erkenntnis*, 44: 49-71.

——— (2002). 'Rethinking Mechanistic Explanation'. *Philosophy of Science* (Suppl.), 69: S342--S353.

Grush, R. (2003). 'In Defense of Some "Cartesian" Assumptions Concerning the Brain and its Operation.' *Biology and Philosophy*, 18: 53-93.

Haugeland, J. (1998). *Having Thought*. Cambridge, MA: Harvard University Press.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: Free Press.

Hitchcock, C. R. (1995). 'Discussion: Salmon on Explanatory Relevance'. *Philosophy of Science*, 62: 304-20.

Kauffman, S. A. (1971). 'Articulation of Parts Explanation in Biology and the Rational Search for Them', in R. C. Buck and R. S. Cohen (eds.), *PSA 1970*. Dordrecht: Reidel.

Kim, J. 2000. 'Making Sense of Downward Causation', in Peter Bogh Andersen, et al. (eds.), *Downward Causation.* Aarhus University Press. 305-21.

Lewis, D. ——— (1983). 'New Work for a Theory of Universals'. *Australasian Journal of Philosophy*, 61: 343-77.

Lycan, W. (1987). *Consciousness*. Cambridge, MA: Bradford Books/MIT Press.

Machamer, P. K., Darden, L. and Craver, C. F. (2000). 'Thinking about Mechanisms'. *Philosophy of Science*, 57: 1-25.

Raichle, M. E. and Mintun, M. A. (2006). 'Brain Work and Brain Imaging'. *Annual Reviews of Neuroscience*, 29:449-76.

Sarkar, S. (1992). 'Models of Reduction and Categories of Reductionism'. *Synthese*, 91:167-94.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

——— (1989). 'Four Decades of Scientific Explanation', in P. Kitcher and W. Salmon (eds.) *Scientific Explanation, Minnesota Studies in the Philosophy of Science XVIII.* Minneapolis: University of Minnesota Press, 3-219.

——— (1994). 'Causality Without Counterfactuals'. *Philosophy of Science*, 61:297-312.

Schaffner, K.F. (1993). *Discovery and explanation in biology and medicine.* Chicago: University of Chicago Press.

Simon, H. (1969). *The Sciences of the Artificial*. Cambridge: MIT Press.

Wimsatt, W. C. (1976). Reductive Explanation: A Functional Account. In E. Sober (ed.), 1984, *Conceptual Issues in Evolutionary Biology*. Cambridge: MIT Press, pp. 369-85.

Wimsatt, W. C. (1994). 'The Ontology of Complex Systems: Levels, Perspectives, and Causal Thickets.' *Canadian Journal of Philosophy* (Suppl.), 20: 207-74.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science (Supplement).* 69: S366-S377.

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