is, arguments with suppressed premises), but by the fact that complete explanations capture all of the relevant causal relations among the components in a mechanism. Understanding the "mechanisms of permeability change" (as Hodgkin and Huxley say), the "mechanistic implications" (as Hille says), and the "underlying biological processes" (as Mauk says) requires understanding the causal structures—the mechanisms—that explain how neurons produce action potentials.

# 3

# Causal Relevance and Manipulation


## Summary

I provide a view of causal relevance that accommodates the mechanistic fragility and historical contingency of neuroscientific generalizations but that nonetheless satisfies constraints (E1)–(E5). I review the limitations of two alternative accounts of causation—Stuart Glennan's mechanical account, and Wesley Salmon and Philip Dowe's transmission account. I use an example from the contemporary neuroscience of learning and memory to defend Woodward's (2002, 2003) view that the causal relevance relations in neural mechanisms are relationships that can potentially be used for the purposes of manipulation and control.


## 1. Introduction

In Chapter 2, I discuss five constraints on explanations in neuroscience, and I argue that any acceptable account of explanation in neuroscience should make sense of their importance. These constraints are:

(E1) mere temporal sequences are not explanatory;
(E2) causes explain effects and not vice versa;
(E3) causally independent effects of common causes do not explain one another;
(E4) causally irrelevant phenomena are not explanatory; and
(E5) causes need not make effects probable to explain them.

These constraints are explained by the fact that successful explanations in neuroscience describe the causal structure of the world. This claim, however, presupposes a view of the causal structure of the world, and one that accommodates (E1)–(E5). In this chapter, I argue for an account of causal relevance that satisfies these constraints. In doing so, I provide an account of causal relevance that makes sense of the norms that scientists use to search for causes and to evaluate causal claims.

My focus on the norms implicit in the practice of neuroscience contrasts with traditional metaphysical projects concerning the nature of causation. First, I do not define "causation" in terms of non-causal concepts, such as "regularity" and "temporal succession." I doubt that any such reductive definition is possible. My colleagues and I have argued elsewhere that at least many cases of causation should be understood in terms of the diverse *activities* that scientists describe in their theories (Machamer et al. 2000; Darden and Craver 2002).[1] Such activities include collision, diffusion, electrostatic attraction and repulsion, gravitation, magnetism, oxidation, and phosphorylation. Activities are no less mysterious than most entities in our best scientific theories (such as atoms, fields, molecules, nuclei, and pituitary glands), and they are no more in need of reductive analysis. From my perspective, causation requires normative regimentation, not metaphysical demystification.

Second, I do not provide an account of the secret connection that Hume sought between a cause and its effect. I will argue that the search for this connection (this cement, glue, spring, or string)—as exemplified by Salmon (1984, 1998) and Dowe's (2000) transmission account and by Glennan's (1996) mechanical account of causal relevance—is sometimes misguided and often distracts philosophers from the aspects of causation that are most important for an account of explanation. This search is sometimes misguided because many causes in neuroscientific explanations are not connected to (that is, in contact with) their effects. For example, in cases of omission and prevention (as when the absence of activity in an inhibitory interneuron allows the post-synaptic cell to fire, or when a competitive antagonist prevents a neurotransmitter from binding to a receptor) there is no hidden connection between the cause and the effect. Such causes work by absences and gaps in connections (or so I will argue). The search for a

[1] This view is further elaborated in Bogen 2004, 2005; Machamer 2004.



connection is distracting because even in cases where one can identify an unbroken connection between a putative cause and an effect, that alone is insufficient to establish that the putative cause is *relevant* to the effect. The search for hidden connections can thus distract one from providing an account of causal relevance, which is much more central to the practice of distinguishing good explanations from bad.

In this chapter, I consider three accounts of causal relations: Glennan's (1996) mechanical account (Section 3), Salmon (1984, 1998) and Dowe's (2000) transmission accounts (Section 4), and Woodward's (2003) manipulationist account (Section 5). The first two—the mechanical and transmission accounts—are each advanced as part of an account of mechanistic explanation, and so it is fair to ask whether they are adequate for that purpose. Each also attempts to identify a hidden connection between causes and their effects. These two views of causation, I believe, are separate paths to the idea that all bona fide causes are found only at the most fundamental ontological levels. All of the real work, one might suppose, is being done by contact action or exchanges of conserved quantities among the most fundamental things. Once one recognizes the limitations of these views of causation, it is much easier to make room for the causal relevance of nonfundamental properties. In the final section, I argue that Woodward's manipulationist view of explanation embodies the standards that neuroscientists (among others) use to discover and evaluate claims about causal relevance, and I show that it satisfies constraints (E1)–(E5). I begin my discussion with the example of Long-Term Potentiation.

## 2. The Mechanism of Long-Term Potentiation

It is widely believed—and there are polls (Stevens 1998)—that brains learn through changes in the strengths of synapses, that is, by changes in the efficiency with which a single action potential in the pre-synaptic cell depolarizes the post-synaptic cell. The most studied form of synaptic plasticity is known as Long-Term Potentiation (LTP). Many believe that LTP, a laboratory phenomenon in which a synapse is strengthened through exposure to a high-frequency pulse, reflects the existence in the synapse of a mechanism for encoding and storing memories. Here, I focus on a type of LTP mediated by a subtype of glutamate receptors that is highly responsive

to the pharmacological agonist, N-methyl-D-aspartate (NMDA). These receptors are called NMDA receptors, and this variety of LTP is commonly called NMDA-receptor dependent LTP. My description of LTP and its mechanisms follows the classic description by Bliss and Collingridge (1993). (For more recent developments, see Squire and Kandel 2000; Lynch 2004; Malenka and Bear 2004.)

A common protocol for inducing LTP involves delivering a tetanus, a high frequency train of stimuli, to populations of pre-synaptic neurons. This stimulus results in a reliable increase in synaptic efficiency. This increase in efficiency is commonly operationalized as: (i) an increase in the slope and amplitude of the excitatory post-synaptic potential in populations of post-synaptic neurons (indicating a larger effect of individual pre-synaptic cells on the post-synaptic response); (ii) an increase in the amplitude of the "population spike" (indicating the synchronous generation of action potentials in the individual post-synaptic cells); and (iii) reduced latency in the population spike (indicating that the post-synaptic action potentials occur faster).

In saying that the tetanus potentiates the synapse, neuroscientists clearly do not mean to assert that whenever one tetanizes a pre-synaptic cell one potentiates the synapse. Nor do they mean to assert that there is a strict law of LTP, in the way that Newton's laws or Ohm's law might reasonably be said to be strict. Of course, one might use the term "law" in a more relaxed sense. It does not matter for my point here. I will show how causal generalizations in biology can function in explanations even if we grant the now well-known reasons for thinking that there are no distinctively biological laws (Beatty 1996; Bechtel and Abrahamsen 2005; Rosenberg 1985; Smart 1963; Weber 2005). Here are four such reasons.

First, LTP is *limited in scope*. It is not a feature of all cells, or of all chemical synapses, or even of all glutamatergic synapses. Its features vary from organism to organism, brain region to brain region, and synapse to synapse. It also varies with developmental stages, with different experimental manipulations, and with the cellular mechanisms used to produce it. In the first full-length report of LTP, Bliss and Lømo (1973) note that the phenomenon varies both across subjects and in the same subject over time. There are several types of LTP, and there are other forms of short- and long-term potentiation that happen at other synapses in the brain. Compared to the genetic code and the theory of evolution by natural



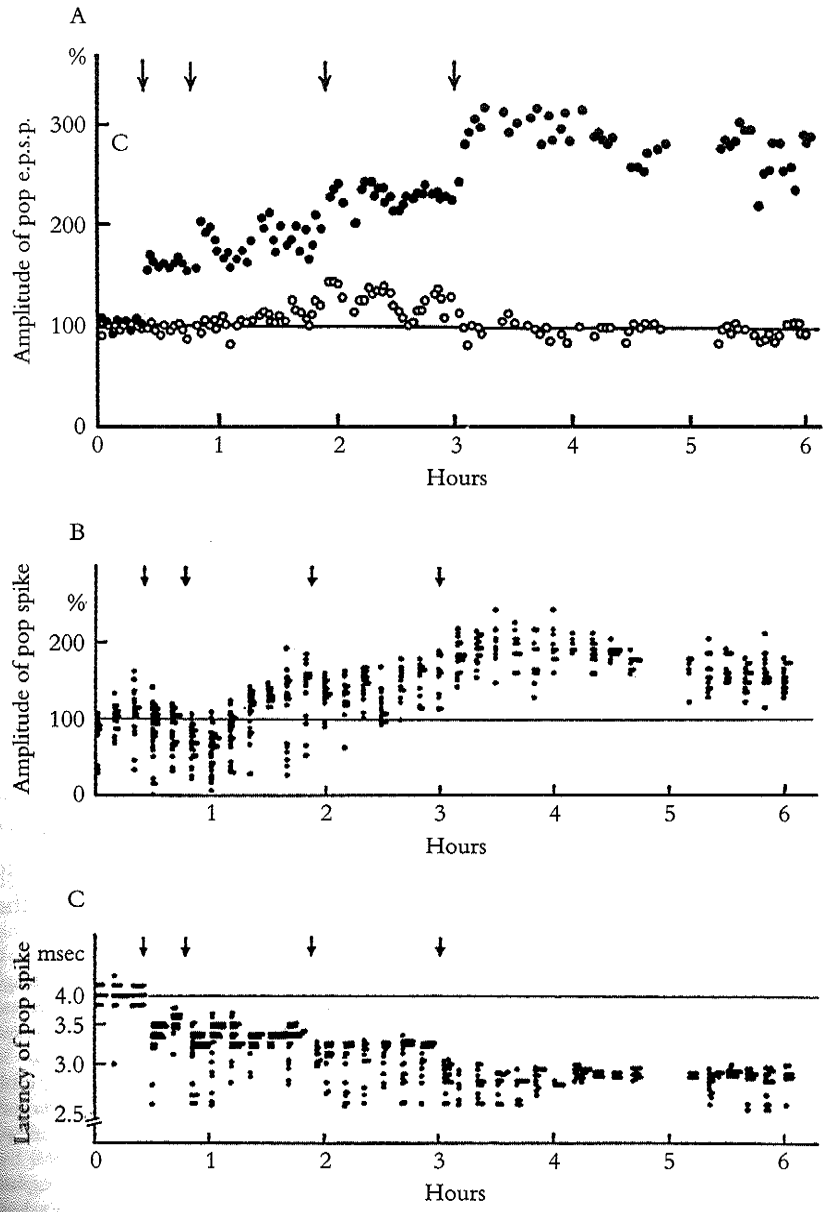


Figure 3.1. Potentiation displayed*

* As (i) increased amplitude of the population e.p.s.p. in A; (ii) the amplitude of the population spike in B; and (iii) the latency of the population spike in C

Source: Compiled and reprinted with permission from Bliss and Lømo (1973: 339–41)

selection, such generalizations as there are to describe LTP have very narrow application.

Second, these generalizations are *stochastic*. Even in those organisms and synapses in which they hold, they hold only some of the time. Bliss and Lømo found all three signs of potentiation ((i)–(iii)) in only 29 percent of their trials. Only feature (ii) appeared in more than 50 percent of the trials. And in only 26 percent of the trials did the synapses show any sign of potentiation thirty minutes after the tetanus. Today, after over thirty years of LTP research, neuroscientists can induce LTP in only roughly 50 percent of their trials.

At least part of the reason that these generalizations are stochastic is that LTP is *mechanistically fragile*.[2] Like most other biological phenomena, LTP varies with features of the stimulus, with background conditions, and with the integrity of the underlying mechanism. By lowering the frequency of the tetanus, one can weaken the synapse rather than strengthening it. By increasing the frequency, one can exhaust or simply incinerate the cell. One can also intervene in myriad ways in the machinery of the pre- and post-synaptic cells such that no potentiation ever occurs. LTP has been reported to vary with such factors as temperature, pH, and time of day (Sanes and Lichtman 1999). If one insists on saying that there are laws of LTP, such laws are at best *ceteris paribus* laws: meaning, roughly, that they hold except when something defeats them, and neuroscientists cannot (now or possibly ever) specify all of the conditions under which they are defeated.[3] Still, neuroscientists know quite a bit about LTP, about the conditions under which it can be induced and maintained, and about the

[2] This point is familiar to philosophers of biology. Rob Wilson, for example, characterizes biology as one of the fragile sciences. This is not the same notion of fragility that Lewis uses to discuss causation in "Causation as Influence" (which has to do with criteria of event individuation), although it is related to other themes in Lewis's paper, especially the discussion of alterations.

[3] There is considerable debate over how to understand *ceteris paribus* laws, about whether they exist, and about whether they can explain anything. As I have formulated the notion of a *ceteris paribus* law, such a law is vacuous either because it is a tautology (the law holds unless it does not) or because we have no idea what it says (the law holds unless X, where X is unspecifiable). The *ceteris paribus* law could also be understood as asserting that the law holds unless there is some factor X that can explain why it does not (Pietrosky and Rey 1995). As Earman and Roberts (1999) point out, this allows there to be a *ceteris paribus* law linking any F to any G, even where F is utterly irrelevant to G. Puffins could act as coincidence detectors in the LTP mechanism if only they were small enough, if they had binding sites for glutamate, if they could change their conformation, and so on, and the fact that they do not meet these requirements explains why the *ceteris paribus* law breaks down. The urge to ground explanatory generalizations in *ceteris paribus* laws has its roots in the idea that one can only explain with strict laws and that *ceteris paribus* laws are hedged strict laws (see Roberts 2004). I do not accept the first of these assumptions. I advocate a different view about how mechanistically fragile generalizations can be explanatory in Section 5.



conditions under which it fails, and neuroscientists count statements about LTP as explanatory.

Finally, the generalizations describing LTP are *historically contingent* (Schaffner 1993a; Beatty 1995; Rosenberg 2001).[4] They are not timeless truths about the brain and its components, but the products of machines cobbled together through evolution by natural selection and soft-constructed in development. The fact that these generalizations are true, in other words, is a contingent product of how life happens to have developed and how a given life happens to develop for each organism. Because of this historical fact, the regularities currently exhibited in biological organisms are not physically necessary, if by that one means that that they could not be different given the laws of physics. There was a time when no organisms in the world exhibited LTP, and there might well be another such time in the future.

These four features of the generalizations describing LTP are not unique to LTP. Instead, they are common to most generalizations in neuroscience and biology generally.[5] They are true, for example, of the causal generalizations describing the mechanism of LTP. Three features make LTP plausible as a potential mechanism of learning: its cooperativity, its associativity, and its input-specificity.[6] Suppose that the experimental set-up involves

[4] The exact sense in which these regularities are contingent is difficult to make precise. Rosenberg argues that, "every regularity in biology will be falsified (or turned into a stipulation) eventually" (2001: 141). This fact about generalizations in biology leads him to claim that there can be no distinctively biological explanations. As he puts it, "One historical fact cannot by itself explain another" (2001: 155; see also Weber 2005: 34). Explanation, on Rosenberg's view, requires the kind of physical necessity found in some of the laws of physics. This conclusion is implausible in the face of the apparent explanatory successes of contemporary neuroscience and biology. I see no reason to believe that one historical fact cannot explain another. They can and they do. Indeed, it is hard to generate examples of explanations that do not explain historical facts by historical facts. Why did the US invade Iraq? Why did the dinosaurs go extinct? Why did AIDS take root among IV drug users? I would accept Rosenberg's conclusion only after exhausting the available options for thinking about how contingent regularities might be explanatory.

[5] Consider Crick's (1988) claim: "Evolution is a tinkerer. It is the resulting complexity that makes biological organisms so hard to unscramble. Biology is thus very different from physics. The basic laws of physics can usually be expressed in exact mathematical form, and they are probably the same throughout the universe. The 'laws' of biology, by contrast, are often only broad generalizations, since they describe rather elaborate mechanisms that natural selection has evolved over billions of years" (p. 5).

[6] There is a sense in which these features of LTP are misleadingly associated with learning. The loose connection is that learning is associative (pairing two co-occurring stimuli, for example), that memories can be primed (cooperativity), and that learning must be specific to associations formed in

stimulating a population of pre-synaptic neurons that converges on a single set of post-synaptic neurons, and that many pre-synaptic neurons converge on the same post-synaptic cells. Depending on whether the pre-synaptic stimulus is weak or strong, it will produce action potentials in a few or in many pre-synaptic neurons. LTP is *cooperative* in the sense that there is a stimulus threshold below which too few pre-synaptic neurons are active to induce LTP. Using a strong stimulus to recruit more pre-synaptic neurons makes LTP more likely at each of the stimulated synapses. LTP is *associative* in the sense that a weak (or sub-threshold) stimulus can produce LTP if it is paired with a strong stimulus to a separate set of pre-synaptic neurons converging on the same post-synaptic cells. Finally, LTP is *input-specific* in the sense that only those synapses that are active during the stimulus are potentiated. These three features point to a common defining mark of LTP: it is induced only when the pre- and the post-synaptic cells are simultaneously active. The synapse thus exhibits a Hebbian form of learning (Hebb 1949).

This defining mark of LTP is explained by a coincidence detector mechanism involving the NMDA receptor (see Figure 3.2). The NMDA receptor gates the diffusion of $Ca^{2+}$ into the post-synaptic cell. When the pre-synaptic neuron is active, it releases the neurotransmitters, glutamate and glycine, which traverse the synapse and bind to receptors on the post-synaptic cell, including NMDA receptors. The NMDA receptors change their conformation to form a $Ca^{2+}$-selective channel through the membrane. If the post-synaptic cell is polarized (that is, resting), the channel is blocked by large $Mg^{2+}$ ions. When the post-synaptic cell depolarizes as a result of activity at non-NMDA receptors (specifically, α-amino-3-hydroxyl-5-methyl-4-isoxasolepropionic acid, or AMPA receptors), the $Mg^{2+}$ ions are repelled from the channels, removing the $Mg^{2+}$ blockage. At this point, $Ca^{2+}$ begins to flow through the channel. The influx of $Ca^{2+}$ and the consequent rise of intracellular $Ca^{2+}$ concentrations then activate a number of intracellular biochemical pathways leading to the changes that constitute a potentiated synapse. In the short term, these pathways add

the environment. Few if any contemporary neuroscientists think that complex associative memories are stored in single synapses, and so it is questionable whether these features of LTP make it directly relevant to learning in the way that these words suggest. The associations relevant to complex forms of learning (for example, semantic memories) are far more likely, given our current understanding of learning, to be formed among distributed representations across populations of neurons than they are to be formed at single synapses.



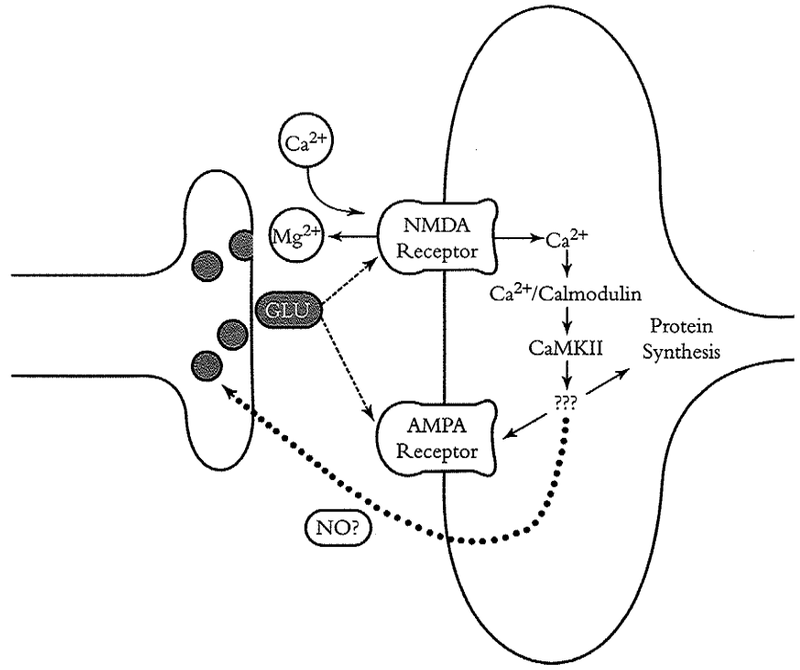


Figure 3.2. A sketch of the synaptic mechanism of LTP*

* Beginning with the release of glutamate (GLU) from the pre-synaptic neuron (left) and terminating with changes to the post-synaptic neuron (right) and/or the pre-synaptic neuron (via nitric oxide, NO)

new receptors to the membrane, or alter their sensitivity to glutamate, or change their $Ca^{2+}$ conductance. Such changes could account for the rapid induction of LTP. In the long term, the biochemical pathways lead to the production of proteins used to alter the structure of the synapse. Some suspect that there is also a pre-synaptic component of this mechanism whereby, for example, the pre-synaptic cell releases more glutamate with each action potential. What matters most for present purposes, however, is that the NMDA receptor gates the induction of LTP. LTP is cooperative because weak inputs do not depolarize the post-synaptic cell sufficiently to remove the $Mg^{2+}$ block. LTP is associative for the same reason: the independent strong input depolarizes the post-synaptic cell sufficiently to remove the $Mg^{2+}$ from the channel. Finally, LTP is input-specific because glutamate opens the NMDA receptors only at the active synapses.

The generalizations describing the causal relationships in this mechanism share all of the features mentioned above for the causal generalization describing LTP. They are *narrow in scope*, holding only for pyramidal cells, or only in NMDA receptors. Many of the causal relationships in this mechanism are stochastic, such as the diffusion of ions or the opening and closing of NMDA receptors. The causal relationships in this mechanism are *mechanistically fragile*. If the concentration of glutamate is too high or too low, if the temperature and pH are not within physiological ranges, or if there is a missing amino acid in the NMDA receptor, many of these causal relationships can break down. Finally, these generalizations are *historically contingent*; before there were NMDA receptors, this mechanism could not work.

Despite the fact that these causal generalizations are limited in scope, stochastic, mechanistically fragile, and historically contingent, they nonetheless describe causal relations in mechanisms that work. They are not mere descriptions of temporal sequences. They relate causes to effects and not vice versa. They do not describe relationships among effects of common causes. And they describe relationships among relevant factors. How must we think about causal generalizations in neuroscience, and the relations that they describe, in order for them to be explanatory despite the fact that they are fragile and contingent? Before getting to my positive view, I first consider two views of causation designed to explicate mechanistic explanation.

## 3. Causation as Transmission

A widespread, if largely implicit, belief about causation is that it involves objects coming into contact and exchanging or transmitting something between them. When the eight ball careens off the two ball into the side pocket, the balls touch and exchange momentum; this exchange constitutes the two ball's causal influence on the eight ball. This view of causation has historical precedent in both science and philosophy, and one might reasonably believe that causal relations in neuroscience (and hence explanations in neuroscience) ultimately are grounded in fundamental causal relations of this sort.

The most influential contemporary expression of this view is found in transmission accounts of causation, especially those of Wesley Salmon



(1984, 1997, 1998) and Phil Dowe (1992, 2000). Salmon developed two transmission accounts: the Mark Transmission account (MT; Salmon 1977, 1984) and the Conserved Quantity account (CQ; suggested by Skyrms 1980; Dowe 1992; elaborated by Salmon 1994 and Dowe 2000). Salmon abandoned MT in favor of CQ, and he ultimately recognized a number of limitations to CQ (Salmon 1989). In each case, the reasons for Salmon's change of opinion show why this view of causation is ultimately unsatisfactory for understanding causal relevance in neuroscience.

The two central constructs of MT and CQ are *causal processes* and *causal interactions*. In this context, processes should *not* be understood as extended events or occurrences, such as production processes or computational processes. Rather, processes are world-lines in Minkowski space–time diagrams; they are things that exhibit consistency of characteristics over time. To introduce three examples that will recur below, a glutamate molecule crossing the synapse, a $Ca^{2+}$ ion entering a cell, and a shadow moving along the ground as a car moves down the highway are all processes. Salmon distinguishes two kinds of process: causal processes and pseudo-processes. According to MT, these are distinguished by the fact that causal processes are capable of transmitting a mark. A mark is a change in some characteristic of a process that occurs when processes intersect. For example, glutamate can be tagged with a radioactive tracer, the NMDA receptor changes its conformation when it binds to glutamate, and a car's shadow is deformed as it passes telephone poles. What distinguishes causal processes (such as glutamate or $Ca^{2+}$ ions) from *pseudo-processes* (such as shadows) is that the causal processes can *transmit* the mark beyond the space–time point at which the processes intersect. A process transmits a mark from space–time point A to space–time point B if and only if the mark appears at each space–time point between A and B in the absence of additional interactions. Once the tracer is attached to the glutamate molecule, additional interactions are not necessary for the glutamate molecule to continue to bear the mark. Likewise, a mark introduced into the car at a local intersection with a pebble (for example, a crack in the windshield) is borne by the car from that point on. The pebble is marked by being broken, compressed, and accelerated. This is not true of the shadow, which is deformed as it passes the telephone pole. The shadow is a pseudo-process because it cannot transmit marks beyond local points of intersection. Two processes (for example, a car's windshield and the pebble)

*interact* causally when they intersect, when both processes are marked, and when the marks are transmitted beyond the point of intersection in the absence of additional interactions. This is an elegant view, indeed.

For both MT and CQ, explaining a phenomenon is a matter of situating it within the causal nexus.[7] As shown in Figure 3.3 (redrawn from Salmon 1984), etiological explanations situate the event to be explained within the causal nexus by tracing the relevant portion of the causal nexus in its past. To say that an event X is part of the explanation for an event Y is to locate X in Y's past light cone and to trace the physical connections—the processes and interactions—linking X to Y.

MT places few restrictions on what constitutes a process or a mark. At least in many cases, neural mechanisms involve parts interacting with one another through contact, and those interactions introduce changes in

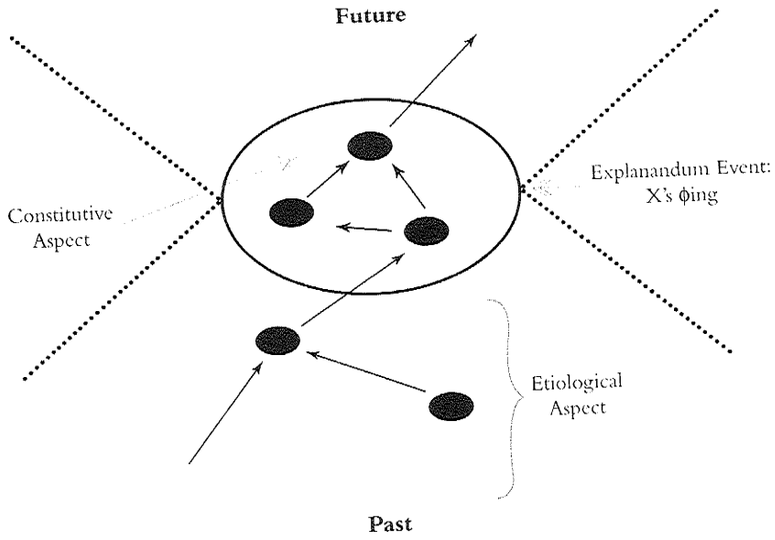


Figure 3.3. Two aspects of causal-mechanical explanation*

* The etiological aspect traces the antecedent causes of the explanandum event (X's $\phi$-ing); the constitutive aspect traces the mechanisms that make up X's $\phi$-ing

*Source*: Adapted from Salmon (1984)

[7] This is a view of etiological explanation. This simple and appealing story does not apply to what Salmon and I call "constitutive" explanations, in which components are not causally (but rather componentially) related to the *explanandum phenomenon*. See Chapter 4.



the properties of the interacting parts. When an enzyme phosphorylates a molecule, and when an electrode injects current, marks are transmitted through local interactions. Despite this prima facie plausibility, neither MT nor CQ can provide a general account of causation in neuroscience.

For example, MT has difficulty accommodating transitory interactions involving marks that are borne only *during* local intersections between processes. Consider the interaction between the glutamate molecule and the NMDA receptor. When glutamate binds to the NMDA receptor, the NMDA receptor changes its conformation and exposes a channel through the membrane. When glutamate detaches from the receptor, the channel returns to its original conformation. The mark is not borne beyond the locus of the interaction. Many of the enzyme-substrate interactions that appear in most molecular explanations in neuroscience have this basic scheme. MT is prevented from accommodating such interactions by its device for distinguishing causal processes from pseudo-processes.[8] Salmon is insistent on this distinction in part because it corresponds to the distinction in special relativity between processes that can be accelerated beyond the speed of light (such as shadows) and those that cannot. Those that can are pseudo-processes, and those that cannot are causal processes. An account of causation in neuroscience, however, need not be bound by a distinction peculiar to physics, especially if the effort to honor this distinction prevents one from accommodating unproblematic kinds of causal relations accepted in neuroscientific explanations.

From Salmon's perspective, however, MT suffers from a more serious limitation. Part of his motivation for developing MT is his desire to treat causation as an empirical phenomenon, open to investigation with the methods of physical science. For Salmon, this involves developing an account of causation free of appeal to counterfactuals (that is, free of statements of the form "if X were to happen, then Y would happen"). Salmon could not see a way to evaluate the truth-values of counterfactuals within his broadly empiricist (though realist) epistemological framework; more specifically, he eschewed efforts to determine their truth-values by

[8] Salmon might respond that the mark fails to persist because of additional interactions inside the channel. Similar counterexamples can arise for fundamental interactions of this sort that cannot be analyzed into interactions among constituent parts. Salmon recognized difficulties accounting for what he calls Y-type interactions that do not involve two processes coming together but rather one process splitting. The channel example involves a Y-type interaction.

considering what happens when counterfactual possibilities (such as X) happen in nearby possible worlds.[9] When he realized that MT makes tacit appeal to counterfactual relations, he abandoned the account (see Salmon 1994). To see how MT makes such tacit appeal, imagine that a shadow (a paradigm pseudo-process) intersects with a glutamate molecule (a causal process) at the same time that the glutamate molecule interacts with a vehicle bearing a radioactive tracer (a causal process). During the intersection with the shadow, the glutamate molecule is marked with a radioactive tracer, and that mark persists beyond the locus of the intersection. The intersection with the shadow, however, is irrelevant to the marking of the glutamate molecule. To rule out cases of this sort, Salmon recognized the need to stipulate that in causal interactions, the mark *would not have been transmitted had the interaction not occurred*. The vehicle satisfies this criterion, but the shadow does not. Salmon's empiricist convictions, however, led him to find this reliance on counterfactuals intolerable.

In "Causality without Counterfactuals" (Salmon 1994), Salmon abandoned MT, citing these concerns about counterfactuals as his primary motivation. He adopted Phil Dowe's (1992) CQ account of causal transmission. According to the CQ account, causal processes and interactions are defined as follows:

> (CQ1) A *causal process* is a world-line of an object that possesses a non-zero amount of a conserved quantity (for example mass, energy, charge).
> (CQ2) A *causal interaction* is an intersection of world-lines that involves exchange of a conserved quantity.[10]

[9] Although Salmon did not express the problem in this way, one can put the problem in terms of the thesis of Humean Supervenience: the idea that there can be no difference in causal facts without a difference in past, present, or future occurrent facts. Roughly, if two worlds are identical with respect to what in fact happens, where, and when, then they cannot differ with respect to what causes what. Failing to conform one's view of causation to this requirement threatens to make causal facts unknowable in principle. If two worlds are identical with respect to their occurrent facts but differ in their causal facts, then no amount of evidence gleaned from the world could ever settle the question of what the causal facts are. Given these epistemic consequences, it seems wise to conform one's account of causation to the principle of Humean Supervenience unless there is a good reason for not doing so. Counterfactual accounts, Salmon might have thought, force one to consider facts other than occurrent facts (for example, facts about nearby possible worlds), thus violating this empiricist loyalty test. Salmon did not develop arguments against counterfactual approaches, such as Lewis (1983), which is intended to be consistent with Humean Supervenience.

[10] It is doubtful that CQ can serve as a reductive account of causation since conservation is an implicitly causal notion. Quantities are conserved if their values remain constant in closed systems, and closed systems are those that have no causal transactions with their environments (see Hitchcock 1995a).



CQ obviates the need for counterfactuals by appeal to conserved quantities that remain constant in the absence of interactions. CQ also presents a view of causation tailor-made for physicalists/fundamentalist metaphysics. If causal interactions are exchanges of conserved quantities, and if conserved quantities are found only at the fundamental level, then all causation is located at the fundamental level.

The added ontological restrictions on processes and interactions in CQ remove the generality that MT has for describing causal interactions in neuroscience. A mark is *any* persistent alteration to a characteristic; so described, marks are ubiquitous. Conserved quantities are not so prevalent. In electrophysiology, as discussed in Chapter 2, explanations appeal to the movement of charges and matter across the membrane. The folding of proteins is similarly described in terms of transitions among stable states through energy-conserving interactions. But the claim that voltage causes NMDA receptors to open makes no explicit appeal exchanges of conserved quantities, and interactions among neurons, such as the LTP phenomenon itself, are not, and do not require for their intelligibility that they be grounded in, conservative interactions.[11] As soon as one begins to talk about causal relations that arise when parts are organized into mechanisms, the transmission view loses traction; its austere descriptive vocabulary no longer applies. When a tetanus induces LTP, there is a causal relationship between an injection of current and the strength of a synapse. Conservation laws do not describe this relationship, and nothing is passed from the tetanus to the strength of the synapse. Although I know of no explanations in neuroscience that violate conservation laws, very few explanations in neuroscience appeal directly to exchanges of conserved quantities. Such exchanges almost always occur well below the "level" of the causal interactions that neuroscientists care about. As a result, very few causal relationships described in neuroscience textbooks can usefully be regimented by assimilating them to conservation laws.

Leaving this descriptive matter aside, however, there are two reasons to doubt that CQ can provide an adequate account of causation in

[11] I neglect here the empirical hypothesis that synaptic changes in the central nervous system are regulated to maintain a constant overall synaptic strength across all of the synapses in a region or system. This could possibly be true. I am merely insisting that its truth is not required for LTP to count as a causal phenomenon.

neuroscience. The first is that transmission theories do not provide a satisfactory account of causal relevance (as required in E4). The second is that transmission theories do not accommodate negative causal relations, such as cases of omission and prevention. MT and CQ can perhaps be supplemented with additional apparatus to remedy these shortcomings. In each case, though, the additional apparatus is an ad hoc adjustment that is untrue to the original motivations for the account and that gives up the simplicity that makes transmission accounts attractive in the first place. A univocal account of causal relations that can do all of the work of CQ and MT without the additional apparatus, such as the manipulationist account I defend in Section 5, looks more promising as an account of the causal relations in neuroscientific explanations.

### 3.1 *Transmission and causal relevance*

Causal relevance cannot be analyzed in terms of exchanges of conserved quantities alone. The causal nexus is a complex reticulum of causal processes and interactions. Only some of them are relevant to any given *explanandum phenomenon*. Providing an etiological explanation involves not merely revealing the causal nexus in the past light cone of the *explanandum phenomenon*. It involves, in addition, selecting the relevant interactions and processes and picking out the relevant features of those processes and interactions (see Hitchcock 1995).

Consider blessed neurons. Suppose our parson electrophysiologist blesses the pre-synaptic neuron with isotonic holy water while delivering a tetanus. The holy water is a causal process transmitting marks and conserved quantities from the micropipette to the neuron. Likewise, the tetanus is induced by injecting current and so involves movement of ions from an electrode into the cell. Matter and energy are conserved in each case. The isotonic holy water is as much a part of the antecedent causal nexus of LTP as is the injection of current. But the blessing is causally irrelevant to LTP.

This example represents a situation at the heart of the search for causes, not just in neuroscience, but generally. The search for causes is not merely a search for what marks what, or what engages in conservative interactions with what, but rather what factors make a difference to the effect. Follow the glutamate molecule from the pre-synaptic cell to the NMDA receptor. The molecule no doubt engages in any number of conservative interactions: it bumps the pre-synaptic membrane; it collides with other molecules; it attracts a passing ion; and it exchanges energy with synaptic enzymes. Each of these interactions involves causal processes exchanging marks or conserved quantities. But only some of these processes and interactions are causally relevant to LTP.



Similar problems arise for causal interactions in which multiple conserved quantities are exchanged. To use an example due to Christopher Hitchcock: a pool cue strikes a cue ball, imparting both momentum and a blue dot of chalk. In the first case, momentum is exchanged. In the last, matter is exchanged. Yet only the first is relevant to the trajectory of the cue ball. We could remove the dot, or change it from blue to green, or change its material constituents in many ways without affecting the trajectory of the ball. Similarly, when an electrophysiologist (ordained or not) lowers the electrode into the cell, the electrode punctures the cell membrane, adds matter to the intracellular fluid, collides with various intracellular molecules, and injects current. Each of these involves an exchange of marks and conserved quantities, but only the current is relevant to LTP. The challenge is to determine which exchanges of conserved quantities are relevant. Because an account of transmission alone does not distinguish relevant from irrelevant markings and exchanges, transmission accounts do not meet condition E4. If explanation is a matter of situating something in the causal nexus, and if the causal nexus contains myriad causally irrelevant processes, features, and interactions, then the explanation includes causally irrelevant features (in violation of E4).

Salmon (1994) acknowledges these problems with his CQ model and admits that they are serious setbacks to his vision of the causal nexus.[12] The problem of explanatory relevance requires a conditional solution. What makes the blessing irrelevant is that the tetanus would strengthen the synapse even if the neuron were not blessed. What makes the blue dot irrelevant is that the imparted momentum would send the cue ball into the eight ball even if the cue did not leave a dot. The goal of an account of causal relevance is to say what makes a difference to what. That goal requires appeal to claims about what would happen or what would be likely to happen if the circumstances were different (or, if it be preferred, to claims about what does or is likely to happen when the circumstances are different; see Bogen 2004).

[12] Dowe (2000) addresses this matter (see his Chapter 7). For criticism of his account, see Hausman (2002); Ehring (2003).

### 3.2 *Omission and prevention*

A second difficulty for transmission accounts concerns the prevalence of explanations that appeal to negative causal factors. This brings me back to the coincidence detector mechanism in LTP. LTP is induced only when both the pre- and the post-synaptic neurons are simultaneously active. As I note above, the crucial features of LTP are explained by the fact that unless the post-synaptic neuron is depolarized when the pre-synaptic neuron is stimulated, $Mg^{2+}$ ions block the NMDA receptor channel and prevent $Ca^{2+}$ from flowing into the post-synaptic neuron. To induce LTP thus requires removing the $Mg^{2+}$ block. Depolarizing the post-synaptic neuron causes $Ca^{2+}$ to enter through the NMDA receptor. But this causal relationship cannot be understood as an exchange of conserved quantities (or transmission of marks) between the depolarization of the pre-synaptic cell and the influx of $Ca^{2+}$.

To see why, focus on the stage in which $Mg^{2+}$ is expelled from the channel in the NMDA receptor. The absence of $Mg^{2+}$ allows $Ca^{2+}$ to enter the cell. The depolarization removes the $Mg^{2+}$ ion from the channel in an interaction that involves exchanging conserved quantities. But no conserved quantities are exchanged between the absence of the $Mg^{2+}$ ion and the influx of $Ca^{2+}$ ions. Absences do not bear or exchange conserved quantities. They are not processes; they are not "things," properly speaking, and they do not exhibit consistency of characteristics over time. Nonetheless, the absence of the $Mg^{2+}$ block does seem to cause $Ca^{2+}$ to enter the cell. At least this is what controlled experiments suggest: when the $Mg^{2+}$ block is in place, the $Ca^{2+}$ does not enter the cell. When the $Mg^{2+}$ block is removed, the $Ca^{2+}$ current begins to flow. In this sense (further restrictions will be added below), the absence of the $Mg^{2+}$ ion makes a difference to intracellular $Ca^{2+}$ concentrations. To the extent that causal relevance is a matter of making a difference, the removal of the $Mg^{2+}$ block, and so the opening of the channel pore, is causally relevant to the induction of LTP.

The example can also be described the other way around. When the $Mg^{2+}$ ion is in the channel, it prevents intracellular $Ca^{2+}$ concentrations from rising despite the fact that glutamate is bound to the NMDA receptor. Does the presence of a $Mg^{2+}$ ion in the channel cause $Ca^{2+}$ not to enter the cell? The $Mg^{2+}$ ion does exchange conserved quantities with $Ca^{2+}$ ions as they enter the channel, but this is not the same as *preventing* an increase in intracellular $Ca^{2+}$. The failure of $Ca^{2+}$ concentrations to rise is not the sort of thing that bears conserved quantities, and it is not the sort of thing with which $Mg^{2+}$ ions can exchange marks. Still, there is no difficulty saying that $Mg^{2+}$ is causally, and so explanatorily, relevant to LTP. The reason is simple: the $Mg^{2+}$ ion *makes a difference* to the $Ca^{2+}$ concentration of the cell, a difference that is revealed, and so can be tested, by removing the $Mg^{2+}$ block.[13]



These examples exhibit two varieties of negative causation.[14] The first, in which the absence of the cause allows an effect (that is, not-C causes E), is commonly called *omission*. The second, in which a cause inhibits or precludes an effect (that is C causes not-E), is commonly called *prevention*. Omission and prevention are common in neuroscience and everyday life. Neurons fire because inhibitory neurons are inhibited. Cells produce proteins because molecules inhibit repressors. Aberrant movements appear in Huntington's disease because of damage to systems that would normally suppress such movements. One does not need to look hard in neuroscientific textbooks to find crucial causal roles for antagonists, blockers, gates, inhibitors, repressors, derepressors, negative feedback, and switches. These are the kinds of systems for which negative causes are crucial.

Jonathan Schaffer (2004) argues, convincingly in my view, that any view of causation that does not include negative causes is sharply at odds with common-sense talk about causes, with scientific judgments about what causes what, and with theoretical applications of the concept of the cause (for example, to understand human agency and moral responsibility). LTP is not unique or in any way exceptional in this respect. Different areas of neuroscience have learned at different rates that they had to include negative causation in their theories for different aspects of brain function. Inhibitory neurons in the brain were not discovered until the middle of the twentieth century. Physiologists and pharmacologists studying the chemical synapse quickly learned that they had to discuss the means by which neurotransmitters are enzymatically inactivated and/or removed from the synapse. The neuroscientists who developed functional brain imaging began

[13] Negative causation also raises problems for Glennan's (1996) mechanical view because absences are not physical parts and cannot "interact," in Glennan's sense of the word, with other parts.

[14] These two can be combined to generate a family of test-problems for views of causation, including cases of double prevention, preemptive prevention, and so on. See Collins et. al. (eds) (2004) for discussions of these and other cases.

by studying only increases in activation during the performance of a task and then realized that they should also systematically investigate reductions in activation as well. Neuroscientists have learned time and again that brain systems can make a big deal out of nothing.

I would add an epistemic point. One discovers and tests negative causal relationships with the same experimental strategies, and negative causal claims are evaluated according to the same normative standards used to evaluate positive claims. If the evidence for testing causal relations is blind to the distinction between negative and positive causes, then our epistemic access to them is no more and no less problematic than it is to positive causes. These considerations make a strong case for accepting negative causes and so place the burden of proof on those who would deny that negative causes exist.

Some philosophers, however, deny that cases of omission and prevention are true cases of causation. There are metaphysical reasons for this view, grounded in the idea that absences, as nothings, have no causal powers.[15] This is a thorny issue, and it is hard to imagine it being resolved decisively. The issue turns in part on how one construes the relata in the causal relations. (Are they events, processes, states of affairs, values of variables, properties, or objects? See Schaffer 2003.) According to the view I recommend, which follows Dretske (1977), Hitchcock (1996), Northcott (forthcoming), and Schaffer (2005), the causal relata are contrasts. For the cause variable, the contrast is between the value of the variable as fixed by the ideal intervention and the value that the variable has in the control condition (that is, without intervention). For the effect variable, the contrast is between the value of the variable in the control condition (when one does not intervene on the cause variable), and its value in the experimental condition (when one does intervene on the cause variable). Causal statements are thus most clearly articulated when they describe a relationship between contrasts: C rather than not-C causes E rather than not-E. Different choices of contrast classes yield different causal claims. To use Dretske's example, it is true that Socrates' ingesting (rather than not ingesting) the hemlock (rather than some non-poisonous beverage) caused him to die (rather than live). It is false that Socrates' ingesting (rather than injecting) the hemlock caused him to die (rather than live). I defend this contrastive view in Chapter 6. Note further that cases of omission and prevention are in many cases merely extremes on a continuum of positive or negative causal relevance. Raising the dose of a drug improves pain relief. Lowering the dose of the drug reduces pain relief. Removing the drug entirely reduces it to a zero-point. Cases of omission and prevention are not outliers in our scientific conception of causation.

[15] This consideration would at best preclude the possibility of omission. It raises no difficulties for prevention, so long as the preventer is a presence.



A second problem raised against the acceptance of negative causes is that there are too many of them, and most negative causes are of no use for understanding explanation in neuroscience. As Dowe (2004) and Beebee (2004) argue, many instances of negative causation run counter to our common sense, scientific, and theoretical uses of the concept of "cause," and no available account of negative causation accepts all and only the intuitively satisfactory instances. If omissions count as causes, then it would appear that I am a cause for all of the things that I might have acted to prevent. Whenever someone spills coffee in someone's lap in a Vienna café, I could have prevented it had I been there and moved the coffee cup, or distracted the waitress, or placed a puffin on the counter, or whatever. I am also the cause of every window's not breaking, for the simple reason that I might have tossed a rock through it. I take very little pride in that fact. It would appear that, in the context of neuroscience, treating cases of omission and prevention as on par with causal processes and interactions makes the project of explaining the brain (that is, discovering its causal structure) much more complicated. The complete etiological explanation (that is, the complete cause) for a phenomenon includes not only all of the factors that actually contributed to its occurrence, but also all of the factors that might have prevented it, no matter how remote. In short, some examples of negative causation are intuitively satisfying and explanatorily salient, and some are not. Such considerations lead many to draw a clear line where they can find it: between positive causation (involving physical connections) and negative causation (not involving physical connections). Dowe (2000, 2004), in particular, argues that the common-sense notion of causation (including cases of negative causation) should be bifurcated into genuine causation, involving exchanges of conserved quantities, and causation*, a counterfactual-laden quasi-causation without connection.

Dowe offers separate accounts of omission and prevention and he shows how these accounts might be extended to cover more complex examples

of negative causation. It suffices for present purposes to examine cases of prevention. On his account, to say that X prevented Y is to say that X caused* not Y. X caused* not Y if:

(P1) X occurred and Y did not, and there occurred an m such that
(P2) there is a causal relation between X and the process due to m, such that either
  (i) X is a causal interaction with the causal process m, or
  (ii) X causes n, a causal interaction with process m, and
(P3) If X (or an alternative preventer) had not occurred, m would have caused Y. (Modified from Dowe 2000, 133–4).[16]

To return to the LTP example, the $Mg^{2+}$ ion is lodged in the channel (X) and the $Ca^{2+}$ does not enter the post-synaptic cell (Y), in accordance with (P1). There is a causal interaction between the $Mg^{2+}$ ion in the channel and $Ca^{2+}$ ions entering the channel (m), in accordance with (P2). If the $Mg^{2+}$ ion had not blocked the channel, then the ion would have moved into the cell, in accordance with (P3). In this way, the $Mg^{2+}$ ion prevented the influx of $Ca^{2+}$ ions, or caused* the ions not to enter the cell.

This account does not so much solve the problem of there being too many negative causes as rename it as a problem for causation*. Dowe recasts the problem in (P1) and (P2), and then appeals to counterfactuals in (P3) to show how the problem can be solved. Dowe does not claim to provide an account of counterfactuals, or a means for distinguishing those generalizations that sustain counterfactuals from those that do not, or a story about how their truth-values are determined. Nor does he provide a means to distinguish those appeals to causation* that are appropriate (the intuitive cases) from those that are not. All he really wants to show is that there are genuine cases of causation that do not require appeal to counterfactuals. The problem of providing an account of the counterfactual in (P3) can then safely be left to those interested in causation*. This strategy is legitimate for Dowe's goal of constructing an empirical account of causation. Dowe's bifurcation effectively banishes omission and prevention from the domain of phenomena over which his theory of causation is required to range. However, my objective is to develop an account of causation that satisfies the norms of explanation in neuroscience. Explanations in neuroscience typically involve instances of both causation and causation*, and so the task of understanding (P3) cannot be left to others. Dowe's account is thus incomplete for present purposes.

[16] Machamer (personal communication) has also suggested an informal version of this view.



The extravagant cases of negative causation can be handled in a number of ways. Some negative causes are too improbable or abnormal to be included in explanatory texts or even counted as causes. Others are ruled out by, for example, legal, moral, and epistemic factors that determine the salience of a fact in a particular discussion (see Beebee 2004). For example, I cannot be held responsible for the coffee spillage in Vienna because I did not know about it (an epistemic claim), I could not reasonably be expected to go to Vienna today for this purpose (a moral claim), and I am under no obligation, legal or otherwise, to prevent the spillage (in the way that perhaps a personal assistant would be). Consider a neuroscientific example: is the gasoline in my car's tank a cause of the instance of LTP in the Petri dish? It is likely true that if I had doused the dish with the gasoline, then the cells would not induce LTP, but it seems odd to think of the absence of gasoline as a cause of LTP. Although I do not have a general formula for ruling out nonexplanatory causes of this sort, it is clear enough that gasoline is neither normally part of cells nor part of their extracellular environment. Gasoline is not part of the set-up or background conditions under which the cell normally operates. It is not a cellular constituent. Gasoline levels do not vary as the mechanism works. The distinction between intuitive and counterintuitive cases is a psychological distinction that is drawn on a number of different grounds in different epistemic contexts.

However, it is a psychological distinction that all parties in this dispute have to accommodate. To remove negative causation from the extension of the term "cause" is only to relocate the problem as a problem for causation*. For the question that naturally arises is: what is the difference between the intuitive and counterintuitive cases of causation*? For the goal of building an account of mechanistic explanation, one cannot simply banish negative causes from consideration. They play too central a role in biological (and neuroscientific) mechanisms. And once they are admitted (either as bona fide causes or as mere causes*), then the extravagance follows automatically.

The problem of causal relevance that I discuss in Section 3.1 and the problem of negative causation that I discuss in this section together present a significant challenge to transmission accounts. For the transmission account

to provide an account of causal relevance, it needs to be supplemented with the idea that causes make a difference to their effects, a difference that can be assessed with controlled experiments. The cue ball, after all, would still have gone in the corner pocket even if it had not been marked with the blue dot. For the transmission account to provide an account of negative causation, it also needs to be supplemented with the idea that causes make a difference to their effects. If the $Mg^{2+}$ ion does not leave the channel, the $Ca^{2+}$ does not enter the cell. But once one has introduced this idea of difference-making (one that many believe requires counterfactuals, such as Dowe's (P3); see Hitchcock 1995; Woodward 2003; see also Bogen 2004) into the account of both positive and negative causation, it is reasonable to ask what further work is left to be done by the requirement that the entire causal chain must involve physical connections or transmission of marks. What justification is there for this further ontological restriction on the notion of "cause"? It is more appropriate to say that even though many cases of causation involve transmission of marks or conserved quantities, this is but one way for something to make a difference to something else. The manipulationist approach that I recommend in Section 5 makes this reliance on difference-making (and the experimental procedures to test it) explicit, shows how this notion should be regimented, and thereby provides a univocal account of positive and negative causal relevance.

## 4. Causation and Mechanical Connection

Let's turn then to Glennan's (1996) mechanical account of causation.[17] His view contains some important insights about how mechanistically fragile and historically contingent generalizations can be explanatory. However, Glennan advances his account as a response to Humean skeptical challenges to causation,[18] and his focus on this classic problem prevents him from developing a normatively adequate account of causation or, consequently, of mechanisms. Glennan has since amended his view (see Glennan 2002) in

[17] I call Glennan's account of causation "mechanical," which should not be associated with my *mechanistic* view of explanation. As my criticisms make clear, I do not think that causation can be explicated in terms of mechanisms, as the mechanical account claims, but I do believe that explanations often describe mechanisms.

[18] I use the term "Humean" to acknowledge debates over Hume's thoughts about causation. I am merely reporting Glennan's rendition of a set of worries traditionally attributed to Hume.



a way that can perhaps handle the objections I raise and in a way that brings his view much closer to my own.[19] Here, I focus only on the earlier view because its limitations bring out some general lessons about the notion of causal relevance in neuroscience.

Glennan (1996) describes Hume's challenge as follows: given a putative cause X and an effect Y, at best, one can observe that X and Y are contiguous, that X precedes Y, and that X-type things and Y-type things are constantly conjoined. One cannot observe the necessity or hidden power by which X causes Y. For Glennan, the challenge is to identify this hidden power connecting X to Y, because that hidden causal power distinguishes cases in which X causes Y from those in which X is merely correlated with or merely precedes Y. Glennan argues that for nonfundamental causes, the hidden causal power is a mechanism linking X to Y: "a relation between two events (other than fundamental physical events) is causal when and only when these events are connected in the appropriate way by a mechanism" (Glennan 1996: 56, 1997). For Glennan, a mechanism is a complex system that produces its behavior by the interaction of a number of parts according to direct causal laws (Glennan 1996: 52). On his account, the tetanus causes the strengthening of the synapse because there is a mechanism (involving glutamate, NMDA receptors, and $Ca^{2+}$) that connects them.

Glennan argues that his mechanical account of causation offers a partial solution to Hume's problem:

> To what degree have we uncovered the secret connexion that binds together causally connected events? At the level of fundamental physics, Hume's problem still remains. We can observe certain regularities, but we cannot offer an explanation of why those regularities obtain. It is not good enough to say that in physics there just are regularities, for there are still questions about which regularities are lawful and causal. Despite the difficulties that remain, we have shown that Hume's problem is not a universal one. In the case of higher-level laws, we can distinguish connections and conjunctions, because we can understand the mechanisms which produce higher level regularities. Very often, the connexion is not so secret after all. (Glennan 1996: 68)

Causation is thus to be understood in two tiers. For nonfundamental causal relations, mechanisms fill the gap between cause and effect with

[19] Indeed, we agree on many aspects of this general framework. My view centers on the same basic ideas and builds from the same philosophical literature. I focus on our differences for purposes of explication.

intermediate causal relations. For fundamental causal relations, there are by definition no mechanisms. While Glennan acknowledges that Hume's problem still arises at the fundamental level, he claims that it is not a problem he needs to confront to understand nonfundamental causes.

Glennan's use of the words "direct causal law" to describe the interactions in a mechanism (Glennan 1996, 1997) has attracted criticism from those who believe that the mechanistic fragility and historical contingency of causation in nonfundamental sciences make talk of universal laws inappropriate (see, for example, Machamer et al. 2000; Woodward 2002; Darden 2002; Glennan 2002). While I agree with the spirit of these criticisms, I also believe that they obscure the progress Glennan makes in thinking about these matters.

Glennan explicitly addresses mechanistic fragility. On his account, nonfundamental causal regularities are sustained by a working mechanism in a range of background and stimulus conditions. Mechanisms (such as the LTP mechanism) break down in inappropriate stimulus conditions, or in abnormal background conditions, or if the components of the mechanism break. Nonfundamental causal regularities are fragile because the mechanisms that sustain them can fail to work.

Although Glennan (1996) does not discuss the historical contingency of nonfundamental causes, the same point applies. Nonfundamental regularities (and the mechanisms that sustain them) are in many cases contingent products of evolution and development. Biological mechanisms are tinkered together (Jacob 1977), and their components are adjusted as variants arise and perish in the course of evolutionary history and as organisms change and develop over their life histories. Such mechanisms have changed considerably over the history of life. They also change over the life of individual organisms. On Glennan's account, such historical contingency subtracts nothing from the ability of a mechanism to act as a hidden connection between present causes and effects. The difference between causal regularities and accidents, for Glennan, is not that causal regularities are timeless and that accidents are historically transitory, but rather that causal regularities are sustained by mechanisms and accidents are not.

To attack Glennan for his use of the term "law" also distracts attention from a serious problem with his attempt to ameliorate the force of Humean causal skepticism, a problem that ultimately ramifies through Glennan's account of causation. The problem derives from a tension between



Glennan's anti-fundamentalism and his attempt to analyze causation in terms of lower-level mechanisms. Glennan states his anti-fundamentalism as follows:

> The mechanical theory of causation rejects a widespread assumption about the nature of causation. I think that it is generally assumed that whatever causal connections are, they ultimately have something to do with the most fundamental physical processes. The closer we are to fundamental physics, the more our statements are about the true causes of things; the further we stray into the higher-level sciences, the more we move away from causal statements and toward mere empirical generalizations. This assumption, however, is what makes Hume's skepticism so devastating. ... Causal statements are typically statements about events regulated by mechanisms, and mechanisms are complex higher-level entities. Only when we talk about interactions governed by fundamental laws does causal talk become problematic. (Glennan 1996: 67)

I believe that mechanists should follow Glennan in resisting causal fundamentalism, but not because such resistance addresses the Humean challenge.

Glennan's anti-fundamentalism does not solve the Humean problem. Although he rejects the view that nonfundamental causal relations are grounded in fundamental metaphysical glue (Glennan 1996: 67), he accepts the weaker intuition that they are grounded in metaphysical glue at lower, yet nonfundamental, levels. For Glennan, the causal relationship between the tetanus and LTP is grounded in causal relations among glutamate, NMDA receptors, and $Ca^{2+}$ ions. Likewise the causal relations between glutamate and NMDA receptors, and NMDA receptors and $Ca^{2+}$ ions, are grounded in still lower-level mechanisms. Glennan's mechanisms are *causal* mechanisms. They are *complex systems* composed of *interacting* parts that *produce* the behavior of the whole according to direct *causal* laws. The italicized words are transparently causal, and the Humean will rightly request an account of these causal terms. If Glennan grounds these causal terms in still lower-level causal mechanisms, then he only staves off ignorance of the nature of causes a little longer. He responds:

> The circularity [or regress] is only apparent. In describing the mechanism that connects the two events I have explained how the events are causally connected. How the parts are connected is a different question. I can try to answer the second question by offering another account of the mechanisms which connect them, but I need not give an account to explain the connection between the events. Indeed such an account would only obscure the causally relevant features of the original

explanation. ... I refer to a mechanism which in turn refers to causal relations, but these latter causal relations are different (and more basic) relations, than the one which I am seeking to explain. (Glennan 1996: 65)

This response is unassailable as a point about *explanation*, but it does not address the worry about *causation*. I agree that one makes progress in explaining LTP by appeal to activities such as binding, changing conformation, diffusing, and phosphorylating. These activities are different from and relatively more basic than LTP. I agree further that it would obscure the understanding of LTP to keep descending through levels of explanation all the way to quarks, strings, and branes. But the Humean problem as stated above is not about explanation. Hume asks, "What is the necessary connexion between cause and effect?" For that question, Glennan's answer is unsatisfying: the mysterious connections at higher levels are grounded in many more and equally mysterious connections at a lower level.

One way out of the regress is to allow it to terminate. One way to do this is to claim that those fundamental causal relations involve exchanges of conserved quantities; I discuss the limitations of that option above. Glennan takes a related approach by claiming that the regress terminates in fundamental laws. He admits that he has no account of how fundamental laws are distinguished from fundamental accidents: "Hume provides a convincing argument that we can have no knowledge of this glue [at fundamental levels], and that talk of such glue may even be unintelligible ... Only when we talk about interactions governed by fundamental laws does causal talk become problematic" (Glennan 1996: 67). Here, Glennan echoes a familiar claim of physicists and many philosophers of physics: belief in fundamental causes is no longer tenable (if it ever was). Many of these physicists and philosophers also argue on these grounds that belief in causes is untenable (or literally false) *tout court* (see, for example, Russell 1913; Norton 2003). This strategy, however, threatens the heart of Glennan's view of causation in a more direct way.

Suppose that one is trying to understand the necessary connection between X and Y (that is, $X \rightarrow Y$) at one level above the fundamental level. Glennan (1996) says that the necessity in the connection between X and Y should be understood in terms of the connections between items at the fundamental level, say, $X \rightarrow a \rightarrow b \rightarrow Y$. Glennan grants that a and b have no necessary connection between them and that talk of such a connection may be unintelligible. But how can a necessary causal connection between X and Y be built out of relations in which there is no necessary connection and for which such talk is unintelligible? The problem then scales up: if the necessary connection between X and Y is problematic, then so are any causal relationships built out of that connection. If the Humean is right about causation at fundamental levels, then when Glennan arrives at the font of causal power at the fundamental level, the well will be dry. In short: the regress terminates or it does not, and either way Glennan fails to solve Hume's problem.



Similar considerations show why Glennan's (1996) mechanical account cannot satisfy (E1)–(E5). Consider (E1), the idea that causal relations are not mere temporal sequences. To meet this constraint, Glennan appeals to the fact that causal relations (as opposed to temporal successions) are underwritten by mechanisms. However, the same worry arises for the causal relations in the mechanism. How are the causal relations in mechanisms distinguished from mere temporal sequences? Glennan answers that the causal relations in mechanisms are interactions governed by direct causal laws. And what are direct causal laws? He answers that they are not mere temporal sequences but necessary connections.[20] Glennan does not explain what it means to say that the direct causal laws are "necessary." It appears that, contrary to the spirit of his mechanical account, Glennan appeals to "direct causal laws" to distinguish causal relations from mere temporal sequences. If so, he does not (and perhaps cannot, given his remarks about fundamental causal laws) develop the resources to adequately distinguish direct causal laws from mere temporal sequences.

A like problem arises concerning the effects of a common cause (E3). Glennan says that "the stipulation that the laws [composing the mechanism] be causal is meant to exclude lawful generalizations which can be explained by common causes" (1996: 55). However, the challenge of (E3) is to find a principled means to distinguish effects of a common cause from causal relations, not merely to stipulate that there is such a difference. Note that there *is* a set of causal relations between the effects of common causes: namely, one that passes via a series of interactions from one effect, through the common cause, to the other effect. Unless Glennan stipulates that the

[20] This response is inconsistent with Glennan's claim, discussed above, that talk of necessity is unintelligible for fundamental causal laws.

bona fide interactions in mechanisms run from causes to effects (that is, stipulates (E2)), there is no way to rule out this set of causal relations as a mechanism linking the two effects of a common cause. Again, the buck of providing an account of causation is passed ever lower in a hierarchy of mechanisms until it is discharged by stipulation in fundamental causal relationships.

Finally, consider the problem of distinguishing relevant from irrelevant causes (E4). Here is a possible description of the LTP mechanism: A glutamate molecule with molecular weight $w$ crosses the synaptic cleft at velocity $v$, collides with a passing protein, alters the position of various amino acids in the NMDA receptor, and lowers the concentration of $Na^+$ in the intracellular fluid. Each step in this bizarre description is true: the molecular weight, the velocity, the collision, the position of the amino acids, and the changes in $Na^+$ concentration each hold for the mechanism producing LTP. This description includes a set of parts and mechanistically explicable interactions. Each stage is linked via a mechanism to its predecessor. Yet no one would claim that this is a good explanation of LTP. This is because the putative explanation is composed of irrelevant features of the synapse. It is not the molecular weight of the glutamate molecule or its velocity that matter, but rather its conformation and charge configuration. It is not the position of a particular amino acid in the glutamate receptor that matters (at least in many cases), but rather the appearance of a pore through the membrane. And it is not the drop in $Na^+$ concentration, but rather the rise in intracellular $Ca^{2+}$ concentration that is relevant to the occurrence of LTP. An account of causation suitable for use in an account of explanation must distinguish causally relevant from causally irrelevant factors. Glennan does not show how this can be done, and so he has not provided a normatively adequate account of the causal relations in mechanisms. As a result, he does not provide a normatively adequate account of mechanisms.

Glennan offers a useful way to explain the mechanistic fragility and historical contingency of neural mechanisms. However, his desire to solve Hume's problem for nonfundamental causes ultimately backfires, driving his account deeper and deeper into a hierarchy of mechanisms. I suspect that many are convinced of the truth of fundamentalism because they endorse a view of causation very much like Glennan's. If this is right, my objections to Glennan's account should help to weaken that motivation. For in Glennan's account, the most pressing questions for a normatively adequate account of causation are stipulated as features of "direct causal laws" at the fundamental level. As a consequence, Glennan neither ameliorates Hume's worries nor satisfies (E1)–(E5). In Section 5, I show how the manipulationist approach satisfies (E1)–(E5) without stipulation and without descending into fundamentalism.



## 5. Manipulation and Causation

I dwell at some length on the mechanical and transmission approaches to causation because each is associated with mechanistic explanation and because each can be used (intentionally or not) to support explanatory fundamentalism. The mechanical view grounds higher-level causal relations in lower-level mechanisms, a grounding process that ends, if ever, only in the most fundamental causal laws. The transmission account is more explicit in its association between causation and properties found only at the most fundamental levels (that is, conserved quantities). The fact that neither of these views provides an adequate account of causation—and in particular, that each struggles to provide an account of causal relevance and negative causation—weakens the attraction of fundamentalism.

To repeat a central theme: causal relevance, explanation, and control are intimately connected with one another. This is particularly true in biomedical sciences, such as neuroscience, that are driven not merely by intellectual curiosity about the structure of the world, but more fundamentally by the desire (and the funding) to cure diseases, to better the human condition, and to make marketable products. The search for causes and explanations is important in part because it provides an understanding of where, and sometimes how, to intervene and change the world for good or for ill. This connection between causation, explanation, and control is also reflected in the procedures that neuroscientists use to test explanations. These tests involve not only revealing correlations among the states of different parts of a mechanism but, further, intervening in the mechanism and showing that one has the ability to change its behavior predictably. More explicitly: to say that one stage of a mechanism is productive of another (as I suggest in Machamer et al. 2000; Craver and Darden 2001), and to say that one item (activity, entity, or property) is relevant to another, is to say, at least in part,

that one has the ability to manipulate one item by intervening to change another. More concretely, to say that LTP is caused by tetanic stimulation is to say that one can potentiate a synapse by tetanizing it.

In embracing this view, I rely closely on James Woodward's account of the role of invariance in explanation (see, especially, Woodward 1997, 2000, 2002, 2003; and Woodward and Hitchcock 2003a, 2003b). Woodward is not especially concerned with neuroscience; however, he is concerned with developing an account of causation adequate for explanations that involve mechanistically fragile and historically contingent generalizations. Woodward (2002) shows how his account of causal relations might be fitted into an account of mechanisms, and Glennan (2002) has followed him in this idea. In this section, I build on that idea by showing how the manipulationist account of causal relevance can satisfy (E1)–(E5). I also show how it can accommodate negative causation.

Woodward's view is currently the most defensible and readable exposition of the manipulationist tradition in thinking about causation both in philosophy (see, for example, Collingwood 1940; von Wright 1971) and in statistics (Cook and Campbell 1979; Freedman 1997). Related ideas appear in Pearl's (2000) notion of a "do operator," the notion of an intervention by Spirtes et al. (1993), and Glymour's (2001) idea of surgically intervening into a causal graph. The central idea is that causal relationships are distinctive in that they are potentially exploitable for the purposes of manipulation and control. More specifically, variable X is causally relevant to variable Y in conditions W if some ideal intervention on X in conditions W changes the value of Y (or the probability distribution over possible values of Y). In the context of a given request for explanation, the relationship between X and Y is explanatory if it is invariant under the conditions (W) that are relevant in that explanatory context. Now I consider the different components of this basic statement.[21]

Woodward construes X and Y as variables, that is, as determinables capable of taking on determinate values. Although this is a common way of speaking in some areas of science and statistics, philosophers have generally

[21] Again, I do not offer this account as a reductive analysis of causation. It would clearly be circular, given that intervention is an ineliminably causal concept. Instead, my account is intended as a necessary condition to be met by relationships of causation and to be explained by any satisfactory metaphysics of causation. Lewis's view of causation, for example, ably captures many of the crucial features of this necessary condition (see Woodward and Hitchcock 2003a for a discussion).



preferred other relata in their accounts of causation. Davidson (1969) and many other philosophers, for example, describe causation as a relationship between events. Salmon (1984, 1998) and Dowe (2000) describe it as a relationship among processes. Others describe it as a relationship among objects, facts, and contrasts. Each of these ways of speaking and thinking about causation can be translated without loss into talk of variables. For example, talk of event and object causation can be translated into talk of a variable that can take on two values {E occurs/is present, E does not occur/is not present}. Talk of causation among processes can be translated by assigning variables to the features of a process or to the magnitudes of the conserved quantities.[22] Similar translations can be made for the other ways of thinking about causation. To view causal relevance as a relationship among variables allows one to consider cases in which the variable may take on any value in a continuum (for example, a dose), to make relative assessments of causal efficacy along that continuum (for example, a dose-response relation), and to consider cases in which there are sharp discontinuities in the effect between one portion of the continuum and another (threshold events, such as action potentials).[23]

The term "intervention" denotes, roughly, a manipulation that changes the value of a variable. It is helpful to think of interventions as well-designed experimental interventions. However, one must not think of manipulations as exclusively the products of human agency. When a stroke damages a brain region, this counts as an intervention on that brain region's functioning. When a meteor strikes the moon, it intervenes in the moon's environment.

The manipulationist view of causal relevance requires that the relationship between X and Y must be *potentially* exploitable for the purposes of manipulating Y in conditions W. One need not actually be able to manipulate X. One might not know how to intervene on X, one might not have the tools, or X might be too small, too big, or too far away for human intervention. Many believe, for example, that a spatial map in the

[22] Those who think of causation as involving activities can make use of the fact that activities have precipitating conditions or enabling properties (that are necessary for or conducive to the occurrence of the activity) and termination conditions or signatures (that is, effects). One can then apply the strategy just described for causation among events, objects, or properties. See Darden and Craver (2002).

[23] As I note above and discuss further in Chapter 6, a contrastive formulation is even more perspicuous. It is a variable X's having one value (rather than some other value) that causes the effect to occur (rather than some alternative). (See Dretske 1977; Hitchcock 1996.)

hippocampus is causally relevant to the ability of rodents to navigate their environments (as argued by O'Keefe and Dostrovsky 1971; O'Keefe and Nadel 1978; Wilson and McNaughton 1993). They believe this in spite of the fact that neuroscientists currently lack the ability to drive a rat through a novel maze by manipulating its spatial map. The ability to do so would no doubt be convincing evidence that the hippocampus is involved in navigation, but this evidence is not required to know that there is a causal relation. What matters is that there is a relationship between X and Y that can possibly be exploited to change Y by changing X, even if no human can or will ever be able to so exploit it. It is a very interesting question how (and how much) we can manage to learn about the causal structure of the world in cases where we cannot intervene in this way. This question is best answered through a detailed look at specific experimental practices in neuroscience. I do not pursue such a detailed investigation in this book (see, for example, Bogen 2001; Bechtel forthcoming). I focus instead on more abstract and general features of the evidence required to establish causal claims.

An *ideal* intervention I on X with respect to Y is a change in the value of X that changes Y, if at all, *only via* the change in X. More specifically, this requirement implies that:

(I1) I does not change Y directly;
(I2) I does not change the value of some causal intermediate S between X and Y except by changing the value of X;
(I3) I is not correlated with some other variable M that is a cause of Y; and
(I4) I acts as a "switch" that controls the value of X irrespective of X's other causes, U. (Adapted from Woodward and Hitchcock 2003a)

These restrictions on ideal interventions are represented graphically in Figure 3.4. Unidirectional arrows represent causal relations, bidirectional dotted arrows represent correlations, and bars across arrows represent a restriction against the represented relation. In this figure, an intervention changes the value of X, surgically removing other causal influences, U, on X (I4). This intervention produces a change in Y that is not mediated directly (I1), by affecting an intermediate variable, S (I2), or by being correlated with some other variable, C, that can change the value of Y



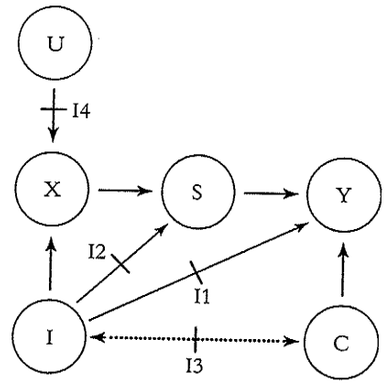


Figure 3.4. An ideal intervention on X with respect to Y*
* Solid arrows represent causal relations; dotted arrow represents correlations; hashes represent the absence of the cause or correlation

(I3). Note that conditions (I1)–(I4) represent the kinds of control that are routinely used and required to test causal and explanatory claims.

The focus on ideal interventions will give rise to objections that experimental situations are often in many ways non-ideal. This is true, and it is an important insight about our epistemic situation with respect to the causal structure of the world. More work remains to be done to say how one can learn about the causal structure of the world if criteria (I1)–(I4) are relaxed, removed, or replaced in order to more accurately describe the complex epistemic situation in which most experimentalists work. The best inroad into that discussion, it seems to me, is to work first on the clear cases and then to see how (and if) the account can be adjusted so that it can regiment non-ideal experimental situations.

Consider the LTP example again. When is it appropriate to assert that a tetanus in the pre-synaptic cell is causally relevant to LTP? One might establish this relationship experimentally by intervening into the pre-synaptic cell, delivering a tetanus, and observing subsequent changes in the strength of the synapse. An experimenter could intervene by injecting current into the cell, by creating an electrical field in a population of pre-synaptic cells, by applying neurotransmitters to the pre-synaptic cell, or by allowing a population of cells to enter its normal burst cycle. What matters is that the intervention makes the pre-synaptic cell fire rapidly and

repeatedly. Suppose that one performs such an intervention and observes a subsequent increase in the strength of the synapse. Such a finding would not warrant belief in the claim that the tetanus is causally relevant to synaptic strength. It is possible that the intervention strengthened the synapse for reasons having nothing to do with the tetanus. Perhaps merely breaking the cell membrane or inserting an electrode into a population of neurons can strengthen synapses (in violation of (I1)). Or perhaps inserting the electrode changes features of neurotransmitter release (in violation of (I2)). Or perhaps one inserts electrodes only when one has the cells in a particular bath solution, and the bath solution strengthens synapses (in violation of (I3)). Or perhaps the injected current is swamped by input from other neurons into the cell (in violation of (I4)). If any of the conditions (I1)–(I4) fails in an experimental protocol, the observed changes in synaptic strength would not be good evidence of a causal relationship between the tetanus and the changes in synaptic strength. When one asserts a causal relevance relation between the firing rate of the pre-synaptic cell and the strength of a synapse, one asserts when one alters the firing rate of the cell in specified ways using an ideal intervention, then one either strengthens the synapse or changes the probability that the synapse would be strengthened.

Each of the activities in the LTP mechanism can be described in the same way. Neuroscientists believe that glutamate opens NMDA receptors because they open when glutamate is applied, but not (or not to the same extent) when isotonic saline or some other neurotransmitter is applied, and not when the binding site for glutamate has been blocked or altered. They are convinced that $Mg^{2+}$ blocks the flow of $Ca^{2+}$ into the post-synaptic cell because they can manipulate $Ca^{2+}$ levels in the cell by changing the concentration of $Mg^{2+}$ or by manipulating the electrical potential that holds $Mg^{2+}$ ions in the NMDA receptor's pore. They are convinced that depolarizing the post-synaptic cell is relevant to the eventual occurrence of LTP because they can keep everything else the same and eliminate LTP simply by clamping the voltage of the post-synaptic neuron at rest. Experiments of this sort show neuroscientists what can manipulate what. On the further assumption that such manipulations are relevantly similar to changes occurring in the brain under the conditions in question, neuroscientists can assume that natural interventions (that is, those not wrought by human hands) produce similar changes in the brain.



### 5.1 *Invariance, fragility, and contingency*

The explanatory generalizations describing these causal relevance relations are stable, or as Woodward says invariant, though not necessarily—or even usually—universal. To say that a generalization is stable is to say that the specified relation between the cause variable and the effect variable holds under a (generally nonuniversal) range of conditions. The conditions under which a generalization might be stable include *stimulus conditions*, *intermediate conditions*, and *background conditions*. Stimulus conditions include conditions explicitly represented as independent variables in the description of the relationship; in the case of X →Y, the stimulus condition is X. The relationship need not be stable across all stimulus conditions. Outside of a normal range of stimulus conditions, the stimulation might have no effect, might weaken the synapse, or might simply damage the cells. The generalization might also be more or less stable under a range of values for the variables intermediate between X and Y, such as $Ca^{2+}$ concentration and $Mg^{2+}$ concentration. Finally, the relationship holds only under a range of background conditions, such as temperature, pH, and available energy. Stable causal relations in neuroscience, in other words, do not hold under all conditions but only under a narrow range of conditions.

The idea that a relationship between variables must be stable to be explanatory is also weaker than the requirement of "contextual unanimity" found in many accounts of causation (for example, Cartwright 1983; Eells 1991; Skyrms 1980). The requirement of contextual unanimity demands, roughly, that if X causes Y, then the relationship between X and Y holds in all contexts. This requirement is too strong for the causal relations in neuroscience precisely because these causal relationships often depend crucially upon the absence of counteracting causes, on the absence of interaction effects, and on background conditions within relatively circumscribed ranges (see Glennan 1997). In contrast to the contextual unanimity requirement, the manipulationist approach allows explanatory generalizations to vary considerably in their stability or invariance and requires only that the generalization should be stable in the conditions relevant to a particular request for explanation (see below).

The fact that generalizations can be more or less stable and still be explanatory is useful for dealing with the fact that causal generalizations in neuroscience are limited in scope, mechanistically fragile, and historically

contingent. Causal relations need not be universal to be explanatory, nor need they be unrestricted in scope, nor need they lack any reference to particulars. All that matters is that there is some stable set of circumstances under which the variables specified in the relation exhibit the kind of manipulable relationship sketched above. Mechanistically fragile generalizations are invariant over a range of values for the stimulus variable, the intermediate variables, and the background conditions. Furthermore, the fact that the relationship is historically contingent, and so in some sense unnecessary, makes no difference to whether the relation is explanatory here and now. $Na^+$ channels produce action potentials today even if no creatures produce action potentials that way 20 million years from now. What matters, again, is that there exists a range of conditions under which one can reliably manipulate the effect variable by intervening to change the cause variable.

Which are the relevant conditions for assessing the stability of a generalization? There is no general answer to that question. Woodward often confines his attention to changes in the values of the variables appearing in the statement of the causal relevance relation. However, this requires at once too much and too little. It requires too much because, as just noted, such relationships might break down under extreme values of the variables appearing in the statement of the relation. It requires too little because, although neuroscientists are often interested in physiologically relevant conditions (that is, the conditions found in intact and healthy organisms), they are just as often interested in disease states in which the stimulus, intermediate, and background conditions are abnormal or pathological. Sometimes they are interested in background conditions well outside the physiological range, as when they try to explain highly contrived experimental effects, to design drugs to interact with the CNS, or to commandeer some part of the CNS for their own purposes. The appropriate range of conditions in which a causal generalization must be stable thus depends crucially upon one's explanatory interests. This does not mean that the causal relations are interest-relative. The causal relevance relations under different ranges of conditions are objective features of the world. However, which of those objective relations is relevant depends on what you are trying to explain.

### 5.2 *Manipulation and criteria for explanation*

According to the manipulationist account, explanatory texts describe relationships between variables that can be exploited to produce, prevent, or alter the *explanandum phenomenon*. Merely being able to manipulate a phenomenon, of course, is not sufficient to explain it. People made babies long before they understood how DNA works. But the wider the range of possible manipulations, and the deeper one's knowledge of how such manipulations change the *explanandum phenomenon*, the more complete is the explanation. As Woodward puts it, a good explanation allows one to answer a range of "what-if-things-had-been-different questions" (w-questions, for short). Deep explanatory texts (or models) provide the resources to answer more questions about how the system will behave under a variety of changes than do shallow explanatory texts. The answers to such questions are evaluated experimentally according to the standards described above.



The manipulationist view readily satisfies criteria (E1)–(E5). Consider mere time-courses (E1). The ability to lay down long-term memories invariably appears after the development of the primary sexual characteristics, but (so far as I know) the latter is explanatorily irrelevant to the former. In contrast, delivery of a rapid and repeated stimulus to the pre-synaptic cell is explanatorily relevant to the entry of $Ca^{2+}$ into the post-synaptic cell. The difference, according to the manipulationist account, is that one could not manipulate the ability to lay down long-term memories by intervening to change the development of primary sexual characteristics (so far as I know), but one can manipulate the tetanus to change the concentration of $Ca^{2+}$ in the post-synaptic cell. This way of dealing with the difference between causation and regular succession has clear advantages over both regularity-based accounts of causation and certain counterfactual views. Both of these alternative views of causation treat at least some cases of regular temporal succession as cases of causation. This is because the values of the two variables, X and Y, are constantly conjoined (*ex hypothesi*) such that whenever the first variable occurs, the second does as well. One could then infer that if X takes a particular value, then Y will take the corresponding value. Nonetheless, it is not the case that one could change Y by intervening to change X. In cases of this sort, the relationship between X and Y supports what Lewis (1979) calls "backtracking counterfactuals," but, as Lewis notes, such counterfactuals are not explanatory. The manipulationist-based approach instead requires causal regularities to fulfill a more demanding requirement, namely that if X is set to x in accordance with (I1)–(I4), then Y will take on the value f(x). This kind of statement

is tested in controlled experiments. Relations that meet this requirement allow one to answer w-questions.

The same strategy can be used to show why causal explanations tend to run from earlier to later (E2). The reason is that, at least in all known cases in neuroscience, one cannot change the past by intervening in current states of affairs. No matter what one does to the pre- or post-synaptic neuron now, one will not change the way that it behaved yesterday. There is no need to *insist* that all causes precede their effects on metaphysical grounds. There are still debates about whether backwards causation is possible in physics (see, for example, Price 1996). Were such a relationship to be demonstrated (using the sorts of ideal experimental manipulations discussed above), one would be justified in asserting that past events can be caused by future events and in asserting that at least in some cases one needs to appeal to future events to explain the past. However, there have been no such demonstrations in neuroscience, and this helps to explain the presumption that explanations in neuroscience are temporally asymmetrical.

Constraint (E3) is that two effects of a common cause do not explain one another in spite of the fact that the occurrence of one allows us to infer the occurrence of the other. Suppose that one pre-synaptic neuron (A) synapses upon two unconnected post-synaptic neurons, $N_1$ and $N_2$; that stimulating A reliably causes $N_1$ and $N_2$ to fire action potentials; and (for simplicity) that $N_1$ and $N_2$ are quiescent in the absence of activity in A. Let X be a variable representing the electrical activity of $N_1$ with the values {firing, not firing} , and let Y be a like-valued variable representing the electrical activity of $N_2$. Under these suppositions, one could reliably infer the value of X from the value of Y and vice versa because $N_1$ and $N_2$ always fire in tandem. That is, there is a robust regularity between X and Y that sustains certain backtracking counterfactuals. Were X to take the value {firing}, then Y would take the value {firing}. And if Y were to take the value {not firing}, then X would take the value {not firing} as well. However, one could not change X's activity by intervening directly to change Y. Nor could one change Y's value by intervening directly to change X. The regularities here do not satisfy requirement W. Examples such as this generalize: if the relationship between two variables is merely a correlation, then one will not be able to manipulate one variable by intervening to change the other. If the two are causally related, then one can manipulate one of them by manipulating the other.



The manipulationist approach also sorts relevant from irrelevant properties and interactions, as required by criterion (E4). (I extend this basic model considerably in Chapter 6 to address issues of nonfundamental causal relevance). To begin with the parson and his micropipette, while it may be true that all pyramidal cells blessed with holy water produce LTP when tetanized, the holy water is irrelevant. One can establish this by intervening in the above sense to remove the blessing, or to change the blessing to a curse, while leaving everything else the same. If one finds that such interventions have no effect on the occurrence (or incidence) of LTP, then one should conclude that the blessing is irrelevant to LTP. Of course, experiments are rarely so clean in the real world. In the history of LTP research, for example, it has been very difficult to determine which of the myriad interactions among intracellular molecules are relevant to the occurrence of LTP (see, for example, Sanes and Lichtman 1999). Part of the reason that these relevance relationships have been so difficult to disentangle is that the intracellular molecular cascades are so complex and causally interwoven that it is difficult to perform the sorts of ideal interventions described above. It is complex, in practice, to determine that one's intervention acts only on the target variable X, and that the intervention changes Y only via X and not through a host of myriad other connections. But these practical difficulties, which are part of what make science challenging and rewarding, do not impugn the overall idea that what one ideally wants to establish is precisely such well-controlled relationships of manipulability.

The final criterion, that the account of causation should allow for improbable effects (E5), requires only a slight modification of the basic argument scheme applied to (E1)–(E4). Many of the causal relationships posited in neuroscience are probabilistic. Tetanizing a pre-synaptic cell produces LTP only 50 percent of the time (with current techniques). If X and Y are only probabilistically related, then any particular intervention to change X might have no effects on Y. As remarked above, what the manipulationist account requires in such cases is that manipulating X changes the probability distribution over possible values for Y. For example, depolarizing the neuron should change the probability that the $Na^+$ channel will open or that the synapse will be potentiated. In neither case is it required that manipulating X makes Y probable (that is, $p(Y \mid X) > 0.5$). The probability of Y might be quite low even under the maximally

effective manipulations of X. Indeed, this matches precisely the way that researchers assess stochastic relationships in neuroscience and elsewhere.

One last point requires emphasis. Nothing in this view of causal relevance makes reference to a privileged level at which all causes act or at which all relevant causes are located. Variables can be fundamental (spin, charm) or nonfundamental (socio-economic status, priming, inflation). All that matters is that they exhibit the patterns of manipulability discussed above.

### 5.3 *Manipulation, omission, and prevention*

A final promising feature of the manipulationist approach to causal relevance is that it accommodates causation by omission and prevention (see Woodward 2002). In cases of omission, such as when the absence of an attractive force allows the $Mg^{2+}$ ion to float out of the NMDA receptor channel, what matters is not the transmission of marks or conserved quantities from the beginning of this mechanism to the end, but rather the fact that one can prevent the $Mg^{2+}$ ion from floating out of the channel by polarizing the cell. Likewise in cases of prevention, such as when the $Mg^{2+}$ ion blocks the channel and thereby prevents $Ca^{2+}$ from entering the cell, what matters is not an exchange of conserved quantities between the $Mg^{2+}$ ion and the non-increase in $Ca^{2+}$ (for there can be no such exchange), but rather the fact that by manipulating the putative cause (positive or negative), one can make a difference in the putative effect (positive or negative).

The ability of the manipulationist account of causal relevance to satisfy (E1)–(E5) and to accommodate cases of negative causation is directly tied to the ability of such generalizations to answer w-questions. This ability provides the kind of rich information about the *explanandum phenomenon* that is typically required of a good explanation. When one knows the relations of manipulability, one can say which interventions make a difference to the *explanandum* and which do not (for example, mere temporal predecessors, temporal successors, irrelevant properties, and the like). In cases where interventions do make a difference, knowing these relations allows one to predict how the *explanandum phenomenon* will be different under a variety of conditions. There is a strong appeal in this connection given that one way to test one's understanding of a phenomenon (as any good test-writer knows) is to test whether someone can say how it will change in novel conditions.



## 6. Conclusion

This view of causal relevance adds an essential normative component to previous accounts of mechanistic explanation. For example, Bechtel and Richardson (1993), like Glennan (1996), argue that mechanistic explanations describe parts and their interactions, but they do not say how to sort interactions from correlations or relevant from irrelevant parts. My co-authors and I (Machamer et al. 2000) describe mechanisms as partly constituted by "activities productive of regular changes," but we do not say what distinguishes productive activities from mere correlations. The manipulationist account clearly makes some progress on this question: X is causally relevant to Y if one can manipulate Y (or, more generally, the probability distribution over values of Y) by intervening ideally on X. X is explanatorily relevant to Y if it is causally relevant.

It is worth noting how much progress can be made in thinking about causation and causal relevance without resolving metaphysical worries about the ultimate nature of causation. The manipulationist approach does not reduce talk of causation to some less problematic notion; the idea of manipulation is causal, and conditions (I1)–(I4) are all stated in causal terms. But it is not clear that a reductive account of causation can provide a satisfactory treatment of causal relevance (that is, one that satisfies (E1)–(E5)). An account of causal relevance should allow one to say which of a number of putative causes actually makes a difference to the effect even if it cannot alone resolve the question of what difference-making *really is*. The diverse examples discussed above (especially cases of omission and prevention) should cast some doubt on the thesis that there is one and only one thing answering to the word "causation." This is one reason why the manipulationist view also remains silent concerning the "hidden connection" between causes and effects. As I have argued, the search for such a connection has led more than one philosopher to develop an account of causation that includes no account of causal relevance. One can complain that the manipulationist account presupposes a metaphysics of causation, and refuse ascent until an account of the metaphysics is provided, or one can recognize the manipulationist account of causal relevance as a normative framework that any adequate metaphysics should satisfy, or better, explain. I do not discuss here whether such metaphysics is required

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