Note to Readers: This paper was written around 2014 for a forthcoming volume of *Minnesota Studies in the Philosophy of Science*. (*Philosophical Perspectives on* C*ausal Reasoning in Biology*, *Volume XXI).* Since publication of that volume has been long delayed and I do not know when it will appear, I am posting this paper now on the Philsci Archive, rather than waiting for it to become even more outdated.

The paper discusses a number of issues having to do with mechanistic explanation in biology. It argues that, as an empirical matter, in mechanistic explanations, there is often a kind of correspondence between difference making information and information about spatio-temporal or geometrical relations of a sort emphasized in causal process theories of causation. This correspondence can function as a constraint on successful mechanistic explanation. The paper also discusses modularity conditions and the circumstances under which they fail. Failures of modularity are distinguished from cases involving redundancy and causal cycles.

**Causation and Mechanisms in Biology**

James Woodward

HPS Pittsburgh

My aim in this essay is to explore some issues concerning mechanistic explanation in biology within a broadly interventionist framework for understanding causation. Along the way, I will touch on some related issues, including the relationship between accounts of causation that focus on the notion of difference-making and those that focus instead on the idea that there must be a connecting process between cause and effect. I will suggest that successful mechanistic explanations often achieve an integration between these two kinds of explanation.

**1. Introduction**

A substantial amount of work in recent philosophy of science has focused on elucidating the notion of mechanistic explanation and the related notion of mechanism. The motivation for this is the recognition that in many areas of science—including the biomedical sciences, but by no means limited to these— the identification of mechanisms is regarded as a major goal of theory-construction: information about mechanisms is thought to be crucial to successful causal explanation, to theory testing, and to the discovery of new causal relationships. I stress this point in order to separate issues about the *importance* of the notions of mechanism and mechanistic explanation in science (where there is general agreement) from issues having to do with how these notions are best elucidated (where there is less consensus).

Notions like “mechanistic explanation” are used in many different ways in science -- sometimes as mere honorifics for theories that are regarded as fundamental, as is perhaps the case with talk of the “mechanism” of natural selection. My goal is to produce an account of a more restricted notion which, very roughly, takes the following form: Suppose system S is characterized by an overall input/output relation *I🡪O*. (For example, ingesting aspirin makes headaches go away, presence of lactose in the environment of E. Coli leads this bacterium to synthesize lactase. ) My target notion of mechanistic explanation explains this *I🡪 O* relationship by identifying (spatio-temporal) parts or components of S and their spatio-temporal organization and patterns of causal dependency among these components (e.g., the state of two components may causally affect the state of a third, which then affects the state of two other components.) A mechanistic explanation shows how, given these dependencies and the organization of the components, the *I🡪0* relationship is produced. This gives one the sense of “opening up the black box” underlying the original *I--> O* relationship.

This general conception of mechanistic explanation is, I believe, shared, at least to a large extent, by many philosophers who have written on the topic (including Bechtel 2006, Craver, 2007, Glennan, 2002, Weber, 2008 and perhaps Machamer, Darden and Craver, 2000) and is also common among biologists, as illustrated by the following remarks:

Mechanism, per se, is an explanatory mode in which we describe what are the parts, how they behave intrinsically, and how those intrinsic behaviors of parts are coupled to each other to produce the behavior of the whole. This common sense definition of mechanism im­plies an inherently hierarchical decomposition; having identified a part with its own intrinsic be­havior, that part may in turn be treated as a whole to be explained (Von Dassow, and Munro, 1999)

One reason for restricting the notion of mechanistic explanation in the way described is to make the topic at least somewhat tractable. In addition, however, I want to separate the project of providing an account of the structure of mechanistic explanation (my topic in this essay) from issues about the *scope* and possible limitations of mechanistic explanation—whether, for example, all or only a limited number of biological phenomena are likely to be fruitfully explained mechanistically. As I see it, the first project is prior to the second -- we need to have an understanding of what mechanistic explanation involves before we can usefully discuss its scope and limits. I leave these latter issues for another occasion[[1]](#footnote-1).

Finally, a word about philosophical methodology: There is some tendency in the literature on mechanisms to argue as though the elucidation of the notion of mechanism is, as it were, an end in itself, with little being said about how this notion connects up with larger questions about causation and explanation. (Perhaps the underlying expectation is that the former project should replace more traditional discussions of causation and explanation.) By contrast, my aim is to explore questions like the following: what is it about the sort of information provided by mechanistic explanations that is especially good or valuable from the point of view of explanation? This requires relating the notion of mechanistic explanation to more general ideas about causation and explanation.

**2. Difference-making versus Connecting Process Accounts of Causation**

As part of the project of placing the notion of mechanistic explanation within a more general framework, I begin with a distinction between two families of “theories” of causation: 1) Difference making *(DM*) theories and, contrasting with these, 2) “connecting process” (*CP*) theories. This distinction is relevant to our topic because a key question about mechanistic explanations is whether they primarily provide DM or CP information or, as I think, elements of both.

The guiding idea of DM theoriesis that causes “make a difference” to their effects, where this is understood in terms of the idea is that causal claims involve a comparison between what happens to the effect in a situation (or situations) in which the cause present and (actual or possible) alternative situations in which the cause is absent or different. (In this sense one might also say DM accounts attempt to capture the idea that effects “depend on” their causes.) The notion of “difference making” (or dependence) may be spelled out in variety of different ways: in terms of changes in the probability of the effect given the presence and absence of the cause, in terms of counterfactuals relating the presence and absence of the cause (and perhaps other actual and potential causes in the situation of interest) to the presence or absence of the effect, and in terms of regularities, as in J.L. Mackie’s (1974) INUS condition account. The interventionist account of causation defended in Woodward, 2003 is one kind of difference-making theory, with the notion of difference-making characterized in terms of what I call interventionist counterfactuals. In the simplest cases in which *C* is a binary variable describing a factor (which can either be present or absent) that deterministically causes *E* (also binary) and there are no complications involving the presence of over-determining or redundant additional causes *C\** which would have caused *E* if *C* had not, the difference-making role of *C* is straightforward: *C* makes a difference when the presence of *C* is followed by *E,* and the absence of *C* is followed by the absence of *E*. When such complications such as overdetermination are present, we can still make sense of the idea that *C* is a difference-maker for *E*, but this requires a more complicated treatment in which one “controls” for other potential causes of *E*. See Woodward, 2003 and section 8 below for more details.

In contrast to DM theories, CP theories attempt to characterize causation in terms of the presence of a “connecting process” between cause and effect. Such connecting processes are often characterized in terms of the presence of characteristic spatio-temporal relationships (e.g., cause and effect may be described as in “spatio-temporal contact”) or in terms of relationships that are viewed as appropriately “mechanical” or “physical”. Examples of the latter are the accounts of Salmon (1984) and Dowe (2000) who take connecting process to involve transfer of conserved quantities such as energy or momentum, as when one billiard ball collides with another. This notion of connecting process appears to have a rather limited usefulness outside of certain contexts in physics. However, several recent accounts of mechanisms in science, such as the well-known theory of Machamer Darden, Craver (2000), as well as an account due to Waskan (2011) also draw on a view of causation and causal explanation that one may think of as broadly in the spirit of CP accounts. These extend the notion of a connecting process in such a way that the presence or operation of an appropriate mechanism between candidate causes and effects counts as such a process. For example, a description of the means by which the electrical signal in a pre-synaptic neuron is converted into a chemical signal in the synapse is taken to be a characterization of a mechanism/connecting process, even though the notion of energy conservation by itself give us little insight into this process[[2]](#footnote-2).

As these illustrations suggest, CP theories apply most directly to so-called *token* causal claims—that is (roughly) claims to the effect that some particular token event has caused another, as opposed to claims that some *type* of event or variable causes another (e.g., action potentials are caused by excitatory post-synaptic potentials). This is a reflection of the fact that it is individual, spatio-temporally located events that are the most obvious candidates for sorts of relata that can be connected by physical processes. There are various ways of extending the CP framework to type causal claims – presumably the most plausible strategy is to regard type causal claims as some sort of generalization over or abstraction from token causal claims -- but to my knowledge no one has worked out the details of this idea.

Again by way of contrast, although it is certainly possible to formulate a DM account of token causal claims, as is illustrated by David Lewis’ well-known theory (1986), DM theories seem to fit type level causal claims or causal generalizations particularly naturally, as a number of the examples discussed below will illustrate.

One of the most distinctive features of CP theories is that they take whether *c* causes *e* to depend just on whether *c* and *e* occur[[3]](#footnote-3) and on whether there is an appropriate connecting process between them. What does or would happen in *other* situations in which *c*  (or a *c*-like event) and *e* (or an *e*-like event) occur or don’t occur (e.g. whether it is true that if c had not occurred, *e* would not have occurred) is taken to be irrelevant to whether there is a connecting process (and hence a causal relationship) between *c* and *e* (Bogen, 2004). By contrast, on a DM conception of causation, information about what does or would happen in situations that are different from the actual situation plays a crucial role since it is central to the idea of difference- making.

DM and CP theories also differ in other ways: for example, they lead to different judgments about whether causal relations are present in cases involving absences. On the simplest version of a CP theory, (*c*) the gardeners failure to water the plants (or, to take a biological example, the failure of some protein to bind to the promoter region of a gene) does not cause (*e*) the death of the plants (or the failure of the gene to be expressed) since in this case there is no connecting process (or physical interaction) from *c* to *e*. By contrast, on a DM theory applied to token causal claims, assuming that the gardener’s non-watering (or the protein’s failure to bind) makes a difference to whether the plants die or the gene is expressed, then these causal claims are true and similarly for the corresponding claims at the type level. The two approaches also reach different conclusions about examples of more biological interest, such as those involving so-called double prevention (or inhibition of an inhibitor) as in the lac-operon model for lactose metabolism (Woodward, 2002).

To further bring out the differences between DM and CP accounts, consider their diverging approaches to the *representation* of causal relationships. Suppose (S1) Billy throws a rock that hits and shatters a bottle. Suppose also there is no back-up cause present that would have shattered the bottle if Billy had not thrown, so that it is true, in S1, that if Billy had not thrown, the bottle would not have shattered. Suppose too that the same is true of a series of situations like S1 in which Billy repeatedly throws at different bottles, so that there is a type-causal relation between his throwing and shattering.

Within a DM approach, we might represent this type-causal relationship between Billy’s throwing and bottles shattering by means of a simple equation:

(2.1) *S= T*.

Here *T* is a variable taking one of two different values, corresponding to whether Billy throws (*T=1)* or not (*T=0*). Similarly, *S* is a variable taking values of 1 or 0 depending on whether the bottle shatters or not. The convention followed in equation (2.1) is that cause variables are written on the right hand side and the effect variable on the left hand side. Note that in this case we have a single equation representing a causal relationship that incorporates information *both* about what will happen if Billy throws *and* if he does not throw—in this way (2.1) conveys difference-making information. A similar point holds if we wish to represent this information by means of a directed graph: an arrow from *T* to *S* can be understood as telling us that the value of *S* depends in some way on the value of *T*, where “depends” implies that the value of *S* will be different for some different values of *T*  -- that is, that *S* will take different values for different values of *T*.

*T🡪S*

Figure 1

Now consider how this information might be represented by a CP theorist. There is no generally established representational device like equations or graphs used in sciences to represent CP information; but a device that might be naturally appropriated for this purpose draws on the so-called neuron diagrams employed by philosophers, as in Lewis, 1986. In S1, when Billy throws, a connecting physical process is present—this being (something like) the trajectory of the rock and the energy/momentum communicated by the rock to the bottle on impact. This connection will be present when and only when the variables *T* and *S* in the previous representation take the value 1. Following the convention adopted in neuron diagrams, we might represent this by means of two shaded nodes connected by an arrow, with the shading indicating that the events of throwing and shattering occur. If, by contrast, we wish to represent situation S2, in which Billy does not throw, we might do so by means of two unshaded nodes, unconnected by arrow, indicating that the throwing and shattering do not occur and there is no connecting process:

*1•🡪 •2*

*1* = Billy throws

*2*= Bottle breaks

where *•🡪 •* means that a connecting process is present

Figure 2a

*not 1o o not 2*

*not 1* = Billy does not throw

*not 2* = bottle does not break

where *o o* means thatno connecting process is present

Figure 2b

The arrows in Figure 1 and 2a have a quite different interpretations. In Figure 1 the arrow does *not* represent a “physical connection” (at least in the sense that CP theorists have in mind) but rather a relation of dependence between T and S. In (2a) the arrow does represent a physical connection and its absence in 2b represents the absence of such a physical connection. In Figure 1 a single unified representation is employed to convey both what happens when Billy throws and when he does not. By contrast, in 2a-2b two different representations are required for this purpose , one describing what happens when Billy throws and one describing what happens when he does not. One issue that will concern us below has to do with the relative utility of these two representations for the depiction of causal relationships in biology. I will suggest that, as this example illustrates, the use of equations and directed graphs to represent causal relationships is often best understood as a matter of conveying DM information along the lines of (2.1) (or an associated graph such as that in Figure 1) rather than just CP information. To the extent that mechanistic explanation makes use of equations, graphs, and similarly interpretable diagrams this is one indication that such explanations draw on DM information.

**3. DM and CP Theories in Relation to Mechanistic Explanation**

How is this contrast between DM and CP theories relevant to issues about the nature of mechanistic explanation? Taken in itself, the information that some system exhibits an *I🡪 O* relationship is naturally thought of as difference-making information: the state or value of *I* makes a difference for the state of *O*. What then does information about the mechanism connecting *I* to *O* add?

It is a common claim of a number of the philosophers (including Bogen, 2004 and Waskan, 2011) that information about underlying mechanisms takes the form of CP information rather than “mere” DM information. This is accompanied by the further idea, more or less explicit, that such CP information is deeper or more explanatorily fundamental than DM information (which is regarded as “shallow”—Waskan, 2011), thus accounting for our sense so that mechanistic information represents an explanatory advance. This line of thought, along with the “actualist” commitments of CP accounts, is nicely expressed in Waskan (2011):

The driving intuition behind some mechanists’ resistance to Woodward’s account of the contents of causal claims is that causal claims make assertions about what *actually* happens rather than about what would happen…. In particular, some contend that causal claims make assertions about actual productive *mechanisms* connecting cause and effect.

In what follows I try to develop an alternative picture of the information provided by mechanistic explanations, according to which such DM information plays a role in such explanations “all the way down” and is not replaced at some deeper level by qualitatively different information about CPs which just have to do with “what actually happens”, as in Waskan’s picture. But at the same time, I will also try to do justice to what I think is correct in the emphasis of Waskan and others on the role of connecting processes and “production” in mechanistic explanation. To do this, I will first sketch the interventionist account of causation on which I will rely and then explain why pure CP accounts of causation (and mechanistic explanation) seem inadequate. This will prepare the way for my positive account of mechanistic explanation. This will take such explanations to be characterized by, in addition to other features, a kind of integration of (or correspondence between) DM features and features having to do with the spatio-temporal relationships on which CP accounts focus.

**4. Interventionism**

Recall that interventionism is one particular version of a DM theory, in which difference-making information is conveyed by “interventionist counterfactuals” describing what would happen were certain interventions to be performed. Such counterfactuals can be used to characterize a very weak and generic notion of type –level causation:

(**M**) Suppose *X* and *Y* are variables. Then, in the simplest case, *X* is a type-level cause of *Y* in some background condition *Bi* (where background conditions are represented by variables distinct from and not causally between *X* and *Y*) if and only if there is a possible intervention on *X* in *Bi* such that if such an intervention were to occur, the value of *Y* or the probability distribution of *Y* would change (cf. Woodward, 2003).

For example, according to (**M**) ingestion of aspirin causes head relief in *Bi*as long as it is true that interventions leading to aspirin ingestion are associated in *Bi* with a change in the incidence of relief among headache sufferers. Similarly, according to (**M**), the level of expression of gene G causes production of protein P if interventions that change the expression of G are associated with changes in the amount of P produced, as might be shown in a genetic knockout experiment.

The intuitive idea underlying (**M)** is that causal relationships are relationships that are potentially exploitable for manipulation and control: *X* causes *Y* if and only if intervening on *X* (in the right circumstances) is a way of changing *Y*. An intervention is an idealized, “surgical” unconfounded experimental manipulation of *X* that changes *X* in such a way that *Y* is affected, if at all, only through this change in *X*. Biological examples of manipulations that are (or may be) intervention-like include gene perturbation experiments, precisely targeted ablation experiments in neurobiology, and randomized drug trials.

(**M)** is a very weak condition: it is satisfied as long as there is *some* intervention on *X* that is associated with changes in *Y* in background *Bi.* The information that (**M)** is satisfied ( hence that *X* causes *Y*) does not say exactly which changes in *X* will be associated with changes in *Y* or about the full range of background conditions in which this will occur. Within an interventionist framework, this more detailed information is provided by more specific interventionist counterfactuals describing more precise patterns of dependence linking *X* and *Y*. As we have already noted, equations are one frequently employed device for conveying such dependency information.

A second notion that is important within the interventionist framework is  *invariance (*also sometimes called stability, robustness, (in)sensitivity) . We may distinguish several different varieties. One has to do with invariance under changes in background conditions. Suppose the relation between *X* and *Y* satisfies **M** for some *Bi* . The invariance of the *X—> Y* relationship under changes in background conditions has to do with whether the relationship would continue to hold in a large or “significant” range of background circumstances different from *Bi* . For example, the relationship between aspirin ingestion and head ache relief is relatively invariant to the extent that relationship holds (or would hold) for subjects in different environment circumstances, in different physiological conditions (within some normal range of course), among populations with different genotypes and so on. Another variety of invariance has to do with stability under a range of changes in the values of the variables *X* and *Y* that figure in the candidate causal relationship. For example, changes in the value of a stimulus (tetanus) *X* might be associated with changes in the firing rate (*Y*) of a neuron for a certain range of values of *X*, but may lead to no response in *Y* outside this range.

Obviously, invariance (in any of its varieties) is a matter of degree and is relative to different ranges of values of variables or sets of background circumstances in the sense that a relationship can be invariant with respect to one such set of background conditions and not with respect to another. Different areas of science and domains of investigation often carry with them different assumptions about which sorts of changes in background circumstances are most relevant or important in assessing invariance. In biology changes that actually occur or are likely to occur on earth, in the past few billion years (as opposed to changes that are merely “possible” in the sense of being consistent with the laws of physics and chemistry) particularly matter for assessments of invariance, but these are not always the only changes that matter, as examples from synthetic biology and bioengineering show.

Invariance can be used to capture one aspect of the idea that at least some causes are *producers* of their effects: causes are producers in the sense that, to the extent that the relation between *X* and *Y* is invariant , then, if *X* were to be introduced in any one of a variety of different circumstances, *Y* would result[[4]](#footnote-4). We will use this idea below in characterizing the role of causal production in mechanisms.

**5. Why Pure CP Accounts are Inadequate**

I claimed above that “pure” CP accounts (that is, accounts that attempt to rely only on CP ideas and dispense with any reliance on difference-making) are inadequate accounts of causation and of mechanistic explanation. The basic reason, illustrated by many examples in the literature, is that the mere existence of a CP between two causal relata fails to give us the kind of fine-grained information about causal relevance (or causal dependence or difference-making) we need for successful causal explanation. When a tennis ball is thrown against a brick wall there is a spatio-temporally continuous connecting process between the movement of the thrower’s arm and the wall, with transfer of energy and momentum, but this fact, by itself, tells us little or nothing about which features of the throw are causally relevant or irrelevant to various properties of the wall, such as its remaining upright or becoming discolored from the dye on the ball. Similarly, consider a glutumate molecule that moves across a synaptic cleft to bind with an NMDA receptor, thus causally influencing the diffusion of Ca++ into the post-synaptic cell. Once again there is a connecting causal process and (we may suppose) transfer of kinetic energy between molecule and receptor but, to a first approximation, this tells us nothing about the features of the molecule that are causally relevant to binding. (Cf. Craver, 2007, p. 92)

Some adherents of CP accounts are unimpressed with such examples, but I see them as pointing to an aspect of pure CP accounts that fits badly with a central feature of explanatory practice. This has to do with the role played by such devices as systems of equations, graphs, and network diagrams. As suggested above, these devices convey information about patterns of difference-making and dependence – information that (as we can also see in the ball- throwing example) is not conveyed just by CP information taken in itself. Mechanistic explanations need to represent such information in some form (whether verbally or more formally) and pure CP accounts seem to lack the resources do this.

To further illustrate this point, consider the following diagrams, due to Eric Davidson et al (2002), representing a regulatory genetic network for endomesoderm specification in the sea urchin embryo.

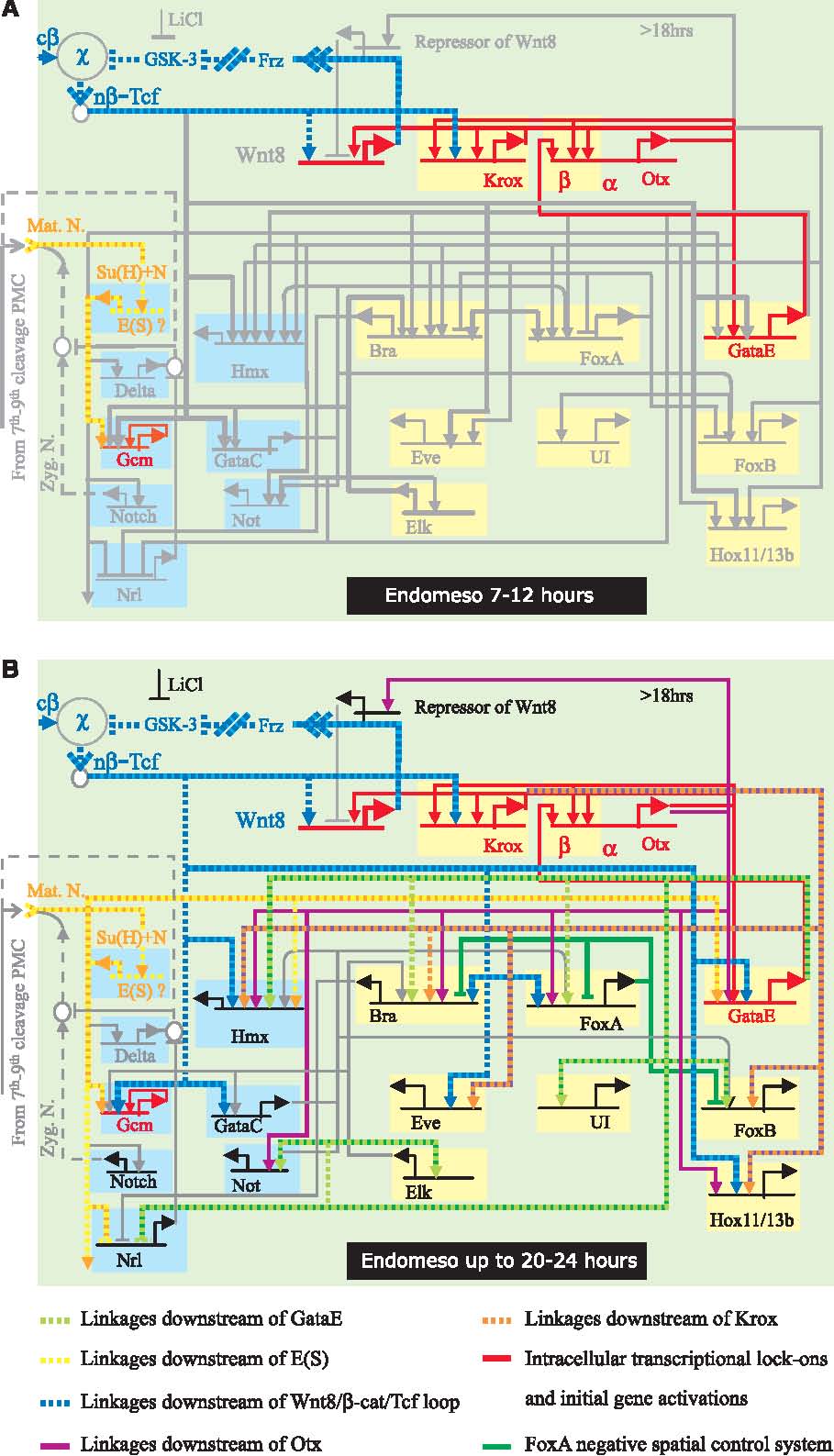


Figure 3

The arrows in these diagrams do *not* represent (or do not *just* represent) “actual” connecting causal processes in the CP sense. Instead these arrows represent relationships of difference- making or dependency— telling us how expression of various genes influences specific targets, that the presence of a certain protein binding to the promoter region for a gene makes a difference for whether the gene is expressed and so on. (This becomes even clearer when we consider that underlying the arrows are “rate equations” specifying the rates or the extent to which various chemical reactions involving genes and proteins will occur (or largely fail to occur) depending on such facts about the interactants as their concentration. These rate equations clearly encode difference-making information.) In this respect, the representational role of the arrows in figure 3 is more like the role of the arrows representing difference-making in Figure 1 than the role of the arrows representing the presence of a physical connection in figure 2a. Indeed, according to Davidson, the causal influences corresponding to the arrows in his diagram can often be represented more precisely by equations or functional relations that are “Boolean” in character— that is, they are representable by functions involving AND, OR, NOT etc.—telling us, e.g. that if both *X* and *Y* are expressed, *Z* will be but not otherwise (*Z=X.Y*). This makes clear that what is being described is a function that tells us both the conditions under which *Z* will be expressed and the conditions under which it will not be expressed-- that is, DM information. Of course it is true that when, say, a transcription factor binds to regulatory region and causally influences gene expression a “connecting process” will be present but this does not mean that the DM information encoded in a gene regulatory network is fully captured just by talk of connecting processes. Rather such processes or their absence reflect or correspond to difference-making relations in a way that I try to elucidate below.

As another illustration, consider the use of the Hodgkin-Huxley (H-H) equations (and their descendants) to model the generation of action potentials. Although some philosophers have suggested otherwise, it is hard to deny that the sort of information about dependency patterns conveyed by these equations plays *some* essential role in the explanation of the generation of the action potential—indeed this was Hodgkin’s and Huxley’s own view, as argued in Woodward, 2017. A full mechanistic explanation of the generation of the action potential includes understanding the molecular details of the processes by which ions pass through pores in membranes and so on, but in my view it would be bizarre to claim that the mathematical analysis embodied in the equations makes no explanatory contribution at all. Again, what these equations convey is dependency or difference-making information. For example, the equation

*I* = *CdVm* /*dt* - *GNa n*4 (m*V* - *ENa* ) – *GK m*3*h*(*V* - *EK* ) + *G*(*V1* –*E1* )

shows how the total current across the cell membrane depends on ( taking the terms on the right hand side of the equation in order) the capacitance current (which in turn depends on the time derivative of the voltage across the membrane), the sodium current, the potassium current, and the leakage current. Additional equations specify how the potassium and sodium currents depends on the gating variables and the voltage. Unless we have some framework, like that provided by the H-H equations, that allows us to see how changes in these factors make a difference for the value of membrane potential over time it seems to me that we do not have an explanation of any depth.

I conclude from these considerations that the idea that successful mechanistic explanations appeal *only* to CP information, and not to any form of DM information is mistaken. I hasten to add, though, that I do *not* conclude that CP accounts of causation are completely wrong-headed. As I see it, the argument of this section supports a role for DM information in the characterization of causation and mechanistic explanation, but not the conclusion that CP information should play no role in the characterization of these notions. Instead, I will explore in the remainder of this essay the suggestion that an adequate account of mechanistic explanation should draw onthe idea that such explanations *integrate* a DM conception of causation with CP-type information. To forestall possible confusion, let me underscore that the conception of causation that I will be working with remains a DM conception; what I will now argue is that, as an empirical matter, in a considerable range of cases, DM information fits with or tracks or corresponds to information about connecting processes and conversely. Mechanistic explanation makes use of this fact.

As a preliminary motivation for this suggestion, consider that despite the apparently deep conceptual differences between DM and CP accounts, adult humans seem to operate in every day life and in science with a conception of causation that brings together central elements of both accounts, rather than leaving them largely or completely independent of one another. To take an obvious illustration, given the CP-type information that there is a connecting process between Billy’s throw and the bottle shattering, we immediately reach the DM conclusion that (barring abnormal circumstances) it was the throw that made the difference for whether the bottle shattered. And if there was no such connecting process (e.g., because Billy’s throw came nowhere near the bottle) but the bottle nonetheless shattered, we would immediately infer that the throw did not make a difference for whether the bottle shattered. This is just one of many examples in which we expect a kind of parallelism or correspondence between DM facts and facts about CPs and associated spatio-temporal relationships. I suggest below that successful mechanistic explanations (at least in many cases) capture or reflect this correspondence.

**6. Sketch of an Account of Mechanistic Explanation**.

With this as background and motivation, I turn to providing a positive account of mechanistic explanation, first sketching some general ideas and then providing illustrations[[5]](#footnote-5). I emphasize that this is proposed as an account of distinctively mechanistic explanations, and not as an account of explanation in general. That is, I assume there can be non-mechanistic causal explanations; one feature of some of these is that they will exhibit DM but not CP information. [[6]](#footnote-6) I also emphasize (as remarked earlier) that the boundaries of the notion of mechanistic explanation are vague (there is no clear pre-analytic notion with sharp boundaries), so that what follows is simply one way (which I hope strikes the reader as somewhat natural or intuitive) of trying to make this notion precise.

**6.1**) In a successful mechanistic explanation, the overall (*I🡪 O*) relationship exhibited by system S is explained by means of a set of generalizations Gi describing causal dependencies among variables describing features of the parts or components of the mechanism as well as information about the spatio-temporal organization of the parts—these dependencies are causal in the sense that they conform to **M** and are invariant at least locally under some range of changes in background conditions. Often the Gi will be more stable, finer-grained, more precise, and have a wider range of application than the original *I🡪O* relationship. Because of this, these generalizations will often be such that they may be used to explain not just the original *I🡪 O* relation but also how the input/output behavior of the mechanism would change if the behavior or configuration of its parts should change. All of this provides us with more information about potential intervention points (and what would happen under such interventions)—thus more information relevant to prediction and control -- than the original *I🡪 O* relationship, thus contributing to our sense that the mechanistic explanation furnishes a deeper explanation.

**6.2)** In thinking about mechanistic explanation, we need to distinguish the role played by components and their spatio-temporal organization from the role played by the generalizations Gi relating the behavior of the components. Typically the Gi will describe dependency relations among features or properties of a number of components, where these features are represented within the interventionist framework by values of variables. The components are literally spatial parts of the system while variables describe causally relevant aspects of those components, and at least in many cases variables will not stand in part/whole relationships.[[7]](#footnote-7) To use an example introduced above, when *X*, *Y* and *Z* are dichotomous variables taking values of 0 and 1, Gi might tell us *Z* = 1 if and only if *X* and *Y* both = 1. As I see it, an account of mechanistic explanation should impose no further requirement that the Gi exhibit any particular functional form such as additivity or linearity. This is desirable since such relationships are the exception in biological contexts.

**6.3** When variables are related in a non-additive way, it obviously will not be the case that the behavior of A (or the “causal contribution” represented by *X*) with respect to component C (or variable *Z*) can be described independently of the value taken by *Y*. Thus when von Dassow and Munro write, in the passage quoted above, that mechanistic explanations ascribe “intrinsic” behavior to components, this should not be interpreted as requiring that each component or variable has a “separable” effect on every other, independently of values taken by other variables, which is what an additivity or no interaction requirement would commit us to. In my view, a more plausible construal of the idea that mechanistic explanations should appeal to generalizations describing intrinsic behavior of components should instead involve the idea that the generalizations Gi (which again may have any functional form whatsoever) describing the dependencies in the mechanism should remain at least somewhat invariant under changes in other causal dependencies holding in the system. This corresponds to a condition that I have elsewhere (Woodward, 2003) called *modularity*, about which more below[[8]](#footnote-8).

Modularity captures, I believe, part of what is defensible in the claim that mechanistic explanations involve a kind of decomposition of some overall *I🡪O* relationship into invariant/stable smaller sets of causal dependencies involving just some of the variables used to characterize the whole system, where these dependencies are “intrinsic” in the sense that they exhibit some degree of independent changeablity. I emphasize again that this understanding of modularity does not require that each spatio-temporally distinct *part* of a mechanism “behave in the same way” independently of what happens to the other parts but rather that *relationships* characterizing interactions among groups of parts remain invariant under changes in other such relationships[[9]](#footnote-9). As we shall see below (Section 7), modularity and the other requirements I impose on mechanistic explanation also allow for causal relationships between relata at different “levels” of organization, as well as “cyclic” causal relationships.

**6.4)** Paralleling this decomposition into modular sets of dependency relations, mechanistic explanation (or at least the kinds of mechanistic explanation I am trying to elucidate) also involves a decomposition of the system of interest into spatio-temporal parts and their relationships. I will have little to say about the criteria guiding this composition, aside from two points. First, a constraint having to do with realism or truth: A successful mechanistic explanation requires that the parts to which it appeals really exist and stand in the spatio-temporal relations described and that the generalizations claimed to describe the dependency relations among the parts be (modulo qualifications resulting from the ubiquitous role of idealization, approximation etc. in science) be true, or nearly so when given the interventionist reading described above[[10]](#footnote-10). Claims about the mechanism of exorcism that appeal to the movement of demons in and out of the soul under the performance of certain ceremonies are not successful mechanistic explanations, even if predictively accurate in some respects. Similarly a purely “functional” decomposition of some *I🡪 O* relationship without any accompanying specification of how such functions are realized materially in spatio-temporal located structures, as is sometimes provided in boxological diagrams in cognitive psychology (sans neurobiology), does not, in my view, provide a mechanistic explanation, whatever its other virtues may be.

A second constraint, already adumbrated, is that the spatio-temporal decomposition must fit with or correspond to the causal generalizations claimed to govern the parts— a notion that I try to say more about immediately below. (Again, I don’t claim that this is a constraint on explanation in general, but rather that it is a constraint on explanations that qualify as mechanistic.)

**6.5)** As already suggested, a crucial feature of mechanistic explanation is that the DM information expressed in generalizations describing the dependency relations among features of the spatio-temporal parts is integrated with information about the spatio-temporal relationships among those parts. “Integrated” in this context means that there are systematic correspondences or connections between, on the one hand, the parts and their spatio- temporal relations and, on the other, DM structure.

There are many possible forms that such correspondences can take. For reasons of space I describe only a few, hoping they will be suggestive.

**(6.5.1)** Relations of causal order or relative causal proximity in the DM sense tend (and as a matter of methodology are expected) to mirror or track spatio-temporal proximity. For example, if *X 🡪 Y* is causally intermediate (or an intervening link[[11]](#footnote-11)) between *W* and *Z* (*W🡪 X🡪Y🡪Z*) when these relations are given a characterization in terms of difference-making, then the parts of the mechanism involved in the *X🡪Y* relationship are typically spatially between (where this may include partial spatial overlap) the parts associated with the *W🡪X* and *Y🡪 Z* relationships, rather than, say, being spatially separated from the *Y—>Z* relationship by the *W—>X* relationship. A similar point holds regarding the temporal betweeness of the steps or processes associated with the *X🡪 Y* relationship with respect to temporal location of the *W🡪 X* and *Y🡪 Z* relations.

**(6.5.2)** Some philosophers suppose it is logically or conceptually possible for *X* and *Y* to stand in a DM relationship R for which there is no process connecting *any* pair of instantiated values[[12]](#footnote-12) of *X* and *Y*. (Following Schaffer 2000, suppose various spells cast by a magician systematically produce different outcomes at arbitrary distances instantaneously). Needless to say, the actual world does not behave like this, even with respect to the DM role of absences[[13]](#footnote-13). Instead, we expect and find that different DM relations are associated with characteristic spatio-temporal patterns or “signatures”, relating at least some instantiated values of these relationships.

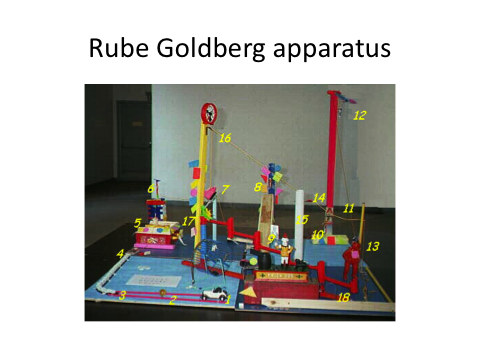
In my view these connections between DM relations and spatio-temporal relations hold (when they do) as an empirical matter in particular cases and not because they are somehow guaranteed to hold apriori—e.g., because of facts about our concept of causation. The connections themselves can take many different forms: Many DM relationships (especially outside of microphysics) are associated with underlying connecting processes which are (at some relevant grain of analysis) spatio-temporally continuous or that at least involve intermediate links or steps which are themselves spatio-temporally localizable. For example, in ordinary circumstances when the impact of a thrown rock causes a bottle to shatter, the rock will move along a spatio-temporally continuous trajectory and will not cause the bottle to shatter unless it comes in spatio-temporal contact with it, in which the case the shattering will occur immediately after impact. By contrast in a case involving transmission of an infectious disease the connecting process involving micro-organisms will not be visible and the onset of the disease will not occur immediately after contact but only after a temporal delay. Thus both causal relationships have spatio-temporal signatures but these are different, depending on the nature of the DM relationship involved.

Moreover, as noted above, intervening/connecting processes/links typically involve variables which are causally between (in a DM sense) the variables related in overall *I🡪 O* relationship. In addition, it is very often the case that when such spatio-temporally connecting processes or localizable intermediate links are found, the variables standing in them are involved in more invariant relationships than the original *I🡪 O* relationship. As a general rule, spatio-temporal proximity and links exhibiting spatio-temporal connectedness tend to go along with greater stability in the DM sense[[14]](#footnote-14) and conversely, so that causes that are producers in the sense of having relatively invariant dependency relations with their effects tend, as an empirical matter, to operate via relatively spatio-temporally continuous processes—i.e., to exhibit “productive continuity”.

Because of these correspondences between DM relationships and spatio-temporal relations, the latter can often be used as an important clue to the discovery of the sorts of detailed and relatively invariant DM relations characteristic of mechanistic explanation, as when one restricts candidates for causal intermediaries in the DM sense to spatio-temporal intermediaries (cf. Darden, 2006). Similarly the absence of appropriate spatio-temporal relationships or connecting processes can be used to exclude certain candidates for causes—as noted earlier, in ordinary circumstances rocks do not shatter windows unless they come in spatio-temporal contact with them.

**7. Some Illustrations**

Consider this machine:



Directions:

1. Turn the handle on a toy cash register to open the drawer.

2. The drawer pushes a golf ball off a platform, into a small blue funnel, and down a ramp.

3. The falling golf ball pulls a string that releases the magic school bus (carrying a picture of Rube Goldberg) down a large blue ramp.

4. Rube's bus hits a rubber ball on a platform, dropping the ball into a large red funnel.

5. The ball lands on a mousetrap (on the orange box) and sets it off.

6. The mousetrap pulls a nail from the yellow stick.

7. The nail allows a weight to drop.

8. The weight pulls a cardboard "cork" from an orange tube.

9. This drops a ball into a cup.

10. The cup tilts a metal scale and raises a wire.

11. The wire releases a ball down a red ramp.

12. The ball falls into a pink paper basket.

13. The basket pulls a string to turn the page of the book!

Figure 4

Suppose initially one is presented only with the information that (*O*) the turning of page of book occurs when only when (*I*) the register handle is turned. This is DM information but it is (at best) explanatorily opaque or shallow; one immediately wants to ask why *I* and *O* are associated in the way described. The information in steps 1-13 above and the accompanying diagram provide a sketch of such an explanation, by exhibiting a set of intervening variables *I, X2, .. X12, O*, associated with different parts of the machine, figuring in a simple causal chain running from *I* to *O*. This gives us information about difference-making relations among each of these variables—e.g., how the impact of the ball makes a difference for whether the mousetrap springs, as in step 5.

The overall handle turning (*I*) -- > page turning (*O*) relation is relatively non-invariant in the sense that many changes either in the environment of the machine or the in any of the intervening links would disrupt this relation. This relation also tells only what will happen under the very specific circumstances in which the register handle is turned and nothing about what will happen under other circumstances. The individual links *I—>X2* etc. are more invariant/stable than the overall *I—> O* link in the sense that there are many changes that will disrupt the latter without disrupting the former: for example, shaking the entire apparatus might set off the mousetrap, thus disrupting the overall *I—> O* relation, but not other individual links such as the handle🡪 drawer opening relation in step 1.

The verbal description and picture above also provide information about the spatial parts of the apparatus and their spatio-temporal organization, with different DM relations being associated with different sets of spatial parts and temporal steps. There is thus a systematic correspondence between this spatio-temporal organization and DM relationships. A causal diagram in the form of a directed graph would represent the DM relationship associated with, e.g., step 8 as causally between steps 7 and 9 and it is also true that step 8 (or the parts of the apparatus associated with the execution of step 8) is spatio-temporally between steps 7 and 9. Finally, the entire mechanism is plausibly regarded as modular in the sense that one can disrupt or change the causal relationships governing some parts or steps without disrupting other causal relationships. For example, one might cut the string in step 3 (thus disrupting the overall *I🡪O* relationship) but this will not disrupt the causal relationship in step 1. Obviously the spatio-temporal separation/localizability of the parts and steps is closely bound up with this independent changeability of the individual causal links making up the mechanism.

Next consider a biological example: the mechanism by which aspirin produces pain relief. Headaches and various other pain-producing disorders cause the production of prostaglandins, which cause swelling of the affected structure and are involved in pain signaling. The synthesis of prostaglandins requires COX enzymes, of which there are two types, COX-1 and COX-2, both operating on arachadonic acids. Aspirin works by inhibiting the synthesizing action of both enzymes, thus preventing the conversion of arachadonic acid to prostaglandins. Both COX enzymes have “tunnels”, through which the arachadonic acid must pass to reach the active sites of the enzymes. Aspirin blocks these tunnels, although the details of how it does so and the active sites themselves differ from COX-1 to COX-2. One reason why these differences are important is that the form of prostaglandin made by COX-1 helps to protect the lining of the stomach; it is the inhibition of this form which accounts for the role of aspirin in upset stomach. Potentially, then, it may be possible to create agents which, unlike aspirin, act only on COX-2 and not on COX-1, thus providing some pain /inflammation relief without causing stomach upset.

This mechanism information seems to fit well into the interventionist framework. It provides more detailed, fine-grained difference-making information (and, relatedly, more detailed, fine-grained possibilities for intervention) that goes well beyond information about the overall difference-making relation between aspirin and pain relief. For example, according to the above description, the presence of prostaglandins makes a difference for whether one experiences pain and inflammation (hence intervening on whether prostaglandins are present in some other way besides aspirin ingestion would also make a difference for pain /inflammation), the synthetic activity of COX-1 and 2 makes a difference for whether prostaglandins are present (so that intervening on this can affect whether there is pain), and whether aspirin is present makes a difference for whether this synthetic activity occurs. Moreover, in assuming that it is possible to intervene to alter the difference- making activities of COX-1 while leaving the relevant difference-making activity of COX-2 intact, we assume that the overall mechanism of action of aspirin is modular at least with respect to these two components. Finally, the biochemical generalizations governing the intervening steps by which aspirin acts, such as the generalizations governing the interaction of the two COX enzymes with arachadonic acid, are far more stable than the overall aspirin 🡪 pain relief relation which they are used to explain.

As a final illustration, recall the regulatory network in Figure 3. Davidson et al. (2002) describe the diagram as providing a “causal explanation” of features of development, adding that “In mechanistic terms, development pro­ceeds as a progression of states of spatially defined regulatory gene expression”. He also writes:

The view taken here is that “under­standing” why a given developmental process occurs as it does requires learning the key in­puts and outputs throughout the genomic regu­latory system that controls the process as it unfolds.

Davidson’s view again seems very close to the picture urged in this essay; the sense in which the diagram provides a mechanistic explanation (or a sketch of an explanation) for a developmental process is that it specifies DM relations among the represented components, how these are organized spatially and temporally and how this contributes to more overall DM relationships.

Figure 3 was constructed by a perturbation analysis in which expression of different genes or signaling pathways was disrupted and the causally downstream effects of such disruption noted. That is, the researchers intervened to alter, e.g., the expression of gene G and then observed what difference this made to the expression of other genes or signaling pathways, with the pathways or arrows in the diagram recording this DM information. (As noted above, the relations represented by the arrows can also be described more precisely by using more specific functional forms). The arrows in the diagram thus communicate information about dependencies that are at least locally invariant under perturbation. Moreover, the use of perturbation analysis requires that the whole network satisfy some modularity-type assumption, according to which it is possible to disrupt individual components or paths without causing a reorganization of the whole network. That is, it is assumed that researchers can intervene to alter expression of a particular gene and that while of course this will alter the downstream causal consequences of the gene, it will not automatically alter the functional relationship between the gene and these consequences. It is assumptions of this sort that permit, as Davidson puts it, “ inferring “normal function from the results of perturbation”. In fact, Davidson himself explicitly describes the regulatory network as consisting of structures that can be thought of as “modules”. As this example illustrates, to the extent that a structure can be successfully represented as modular, this greatly facilitates the task of learning about the mechanism through experimental interventions on its individual parts.

**7. A More Detailed Look at Modularity**

Although, for reasons just described, mechanistic representations or explanations reflecting decomposition into modular structures have many virtues, there is considerable controversy about just what the assumption of “modularity” commits us to and about the extent to which various biological systems are or may be usefully represented as modular. One initial question concerns—to put matters crudely—whether modularity should be thought of primarily as a feature of systems as they exist in nature or instead as a feature of our *representations* or *explanations* of such systems[[15]](#footnote-15). My own preference is to think of modularity in the first instance as a feature of representations or explanations. This allows for the possibility that a system might have several different candidate explanations, perhaps at different levels of grain or resolution (including spatial or temporal resolution) , some of which are modular and others or not. We can then go on to say that a system itself is modular or not , depending on whether it has *some* modular representation, perhaps at some level of analysis at which such a representation is viewed as fruitful[[16]](#footnote-16). Degrees of modularity can be handled similarly.

As a point of departure, let me underscore again what the modularity amounts to in the present context. Modularity has to do with whether various causal *relationships,* as described in some representation/explanation, remain stable (continue to hold) as other causal *relationships* as describedin the explanation are changed. Suppose that that the system of interest is represented as governed by two equations describing distinct causal relationships:

1. *X2 = 2X1*

and

1. *X3=3X2*,

and that *X1=1*, so that *X2 =2* and *X3 =6*. Suppose that an intervention occurs that sets *X2 = 3* . This disrupts the causal relationship (i), since the value of *X*2 is no longer determined by *X1* according to (i). This intervention will also change the input into equation (ii) and if (ii) continues to hold, will lead to a value of *X3=9* different from the original value *X3=6*. This does *not* reflect a failure of modularity. Instead, the system as represented behaves in a modular way, since (ii) continues to hold under changes in (i). A failure of modularity would occur if the relationship (ii) would also change under this intervention that disrupts (i). In other words, what matters for a failure of modularity is not whether downstream *outputs* change when upstream causal relationships are disrupted or upstream inputs are changed (this will typically happen in modular systems, as long as they lack redundancy—see below), but whether other causal *relationships* (besides the relationship disrupted) change under this disruption. Put slightly differently, a failure of modularity occurs when a structure of causal relations is represented as reorganizing globally under what is represented as a local change altering causal relationships among just some of its components.

Ralph Greenspan (2001) gives an illustration of this possibility (Figure 5) involving a hypothetical genetic regulatory network, which is also discussed in Mitchell (2009). (Nodes in the network represent genes and the paths between nodes represent regulatory dependencies of various sorts among the genes.)

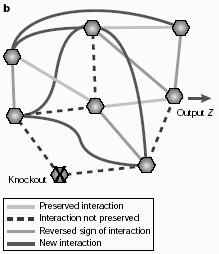
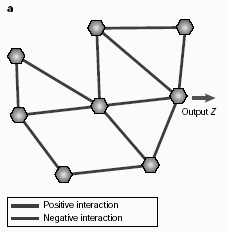


Figure 5

The diagram in (5a) represents the form originally taken by the network. An intervention on one node/gene in this network (in the form of a knock-out experiment) leads to a change in the organization of the whole network, as represented in (5b).

Along with Mitchell, I take this to be a case in which the representation in (5a) hand side fails to be modular. One might imagine (at least) two different reactions to this fact, depending on whether one thinks that a failure of modularity at any accessible level of analysis is a real possibility or whether one thinks it merely indicates that one has failed to correctly represent the causal structure of the system[[17]](#footnote-17).

(**7.1**) Reaction 1: Not only is the representation (5a) above non-modular (as shown by what happens in 5b), it is entirely possible that there is *no* accurate representation of the system that is modular, at least at any level of analysis at which researchers wish to work. All candidate explanations may be such that they exhibit global reorganization under local interventions. It is an empirical question whether this happens and, assuming that to qualify as “mechanical” an explanation must meet some modularity requirement, it is also an empirical question to what extent various biological systems are susceptible of mechanical explanation.

**(7.2)** Reaction 2: When a candidate representation/explanation shows an apparent failure of modularity, this indicates that the representation is defective or incomplete in some way. In particular, if a genetic network appears to reorganize in the manner described by Greenspan, that indicates the original specification of the network was (causally) incomplete. After all, (one might argue) there must be some causal explanation for why the network reorganizes as it does when one of the nodes is perturbed. Obviously, this information is not represented in diagram (5a). . A fully adequate causal explanation should accurately represent how each node would change under any possible intervention on any other node and (one might suppose) such an explanation will necessarily be modular. On this understanding, one might think of modularity as an explanatory or representational ideal: One should keep refining candidate explanations until the system represented appears to behave in a modular fashion.

I don’t want to disparage the idea of looking for a modular representation when a system behaves in an apparently non-modular way (according to the representation we are currently using)—this will often be worth doing. However, the claim that failures of modularity are always merely apparent and always can be dealt with by replacing a non-modular representation by a modular one seems too facile. Some interventions on a variable *X* do seem to result in a “re-organization” of causal relations elsewhere (not involving variable *X*) in a system. One reason for thinking this is that, as observed above, there are often additional constraints that we think, as an empirical matter, causal relationships must satisfy in addition to DM constraints. In particular, as argued above, spatial and temporal relationships, appropriately integrated with DM relationships, are one important source of such constraints. Consider a brain in which at time t there is no no neural projection (and thus nothing that might serve as a basis for a relevant physical connection) from neural structure *X* to *Y* and assume also that without such a projection, *X* cannot causally influence (make a difference for) *Y*. After a change in some distinct neural area *Z (*perhaps as a result of an injury*),* the brain reorganizes at a later time t’ such that a functional neural connection from *X* to *Y* now comes to exist, with the result that changes in *X* now lead to changes in *Y*. This seems like a case in which the causal relationship between *X* and *Y* changes, as a result of the change in *Z* and hence in which there is a genuine failure of modularity at least at the time scale associated with the growth of the new neural connection. In this case, the integration of DM and spatio-temporal or connecting process information helps us to sort out when we have genuine failures of modularity, because there are independent constraints having to do with spatial connections on whether DM relations can be operative. That is, the absence of an anatomical connection (and thus the possibility of a connecting process) between *X* and *Y* at t and the presence of such a connection at t’ leads us to think that there has been a genuine change in the DM relation between *X* and *Y*, rather than the manifestation at t’ of a relationship that was “there” all along. Put differently, attention to the integration between DM and spatio-temporal information gives teeth to the modularity requirement in the sense that it describes circumstances in which the requirement would be false.

These remarks also have implications for the intervention requirement (**M**). In particular they suggest that if (**M**) is to be useful there must be some tacit restrictions on what counts as “background circumstances” in (**M**). If we say that neural area *X* causally influences neural area *Y* as long as there are some background circumstances in which interventions on *X* will change *Y* where such background circumstances include all possible arbitrary changes elsewhere in the brain, we will probably end up drawing causal arrow from every neural area to every other, producing a causal diagram that is useless. (Every neural area can influence every other under some background circumstance corresponding to some possible change elsewhere in the brain.) Instead we tacitly restrict the relevant background circumstances in various ways—e.g., perhaps to those that are consistent with the anatomical connections obtaining at present in some actual brain of interest.

As an additional illustration, suppose the various parts of the apparatus in figure 4 are disassembled, and lie in a disorganized heap. Suppose it is argued that it is nonetheless true that pulling the cash register handle causes turning of the pages of the book on the grounds that there are some background circumstances (namely those in which the apparatus is assembled as in Figure 4) in which such pullings make a difference for whether page turnings occur. An obvious response is that there is an important difference between what would happen in some set of merely possible circumstances in which the apparatus is assembled in a certain way (given that it is now a disorganized heap) and what would happen given the actual arrangement of those parts. We shouldn’t draw an arrow from *X* to *Y* merely because it is possible for *X* and *Y* to be connected up in such a way that interventions on *X* will be associated with changes in *Y*—or at least we should recognize that if we do this, we are likely to end up with a not very useful explanation/representation in which nearly every variable is represented as connected to every other (since e. g., in the case of the disassembled apparatus each of the parts might be capable of being connected to any other part in a way that causally influences it). Again it seems natural to suppose that when we connect up the parts in the way represented in figure 4, we create new arrows or causal relationships that did not previously exist.

I conclude that, despite some incautious remarks of mine to the contrary in an earlier paper (Hausman and Woodward, 1999), Greenspan and Mitchell are correct to think it possible for there to be biological systems in which modularity fails. This of course leaves open the additional question of how wide-spread failures of modularity are and to what extent it is good strategy, when presented with apparent failures, to look for alternative representation/explanations that come closer to satisfying modularity constraints. I will also add that in my view it is a mistake to conclude, that because modularity sometimes fails, it is not an interesting or important feature of biological systems or explanations of such systems when present. It is a bad habit of philosophers to suppose that if feature X does not hold universally, in all circumstances, it is not worth paying attention to.

**8. Modularity, Robustness, Redundancy, Downward Causation, and Cycles.**

Careful biological discussion distinguishes between the kind of network reorganization considered by Greenspan above, which is sometimes described in terms of *robustness,*  and genetic *redundancy.* Robustness arguably reflects a failure of modularity but redundancy does not. In redundancy, a gene *G1* causally contributes to some trait *P*, but *P* remains when *G1* is knocked out because some other gene *G2* becomes active and replaces the causal contribution of *G1* to P. Cases of this sort are broadly similar in structure to some of the cases of linked overdetermination or redundant causation discussed in the philosophical literature. They may be represented graphically as follows:

G1 G2

P

with the parameterization that *P* is “on’ when either *G1* or *G2* is on and not otherwise, *G2* is off if *G1*is on and *G2* is on if *G1* is off, and initially *G1* is on and *G2* is off, with *G1* then being turned off, so that at that point *G2* becomes on.

Following Woodward, 2003, we can capture the idea that initially *G1* is the “actual cause” of *P* in terms of the following: if *G2* is fixed at its actual value (= off), intervening on *G1* will change the value of *P*, so that in this sense *G1* is an actual cause (or actual difference-maker) for P. On the other hand, fixing *G1* at its actual value = on, intervening to change the value of *G2* from its actual value=off to on to the value = on makes no difference to the value of *P*, which remains at the on value. Thus, initially *G2* is not an actual cause for *P*. But once *G1* is changed to its “off” value, the value of *G2* becomes “on”, and parallel reasoning shows that *G2* qualifies as the actual cause of *P*. In this way, interventionist or difference-making ideas can be extended in a natural way to cases in which redundant causation is present—in other words, causes in systems in which redundancy is present still may be regarded as difference-makers, but we need to appeal to what would happen under multiple interventions to bring out their difference-making role. It is also worth adding that this treatment maps straightforwardly onto what might be done experimentally to disentangle the causal structure of a system with redundancy: one carries out a series of multiple intervention experiments in which one intervenes on both *G1* and *G2* simultaneously, fixing one at some value and changing the value of the other.

This treatment also makes it clear that with redundancy of this sort , there needn’t be any failure of modularity. Knocking out *G1* does not change the relationship between *G1* and *G2* or between *G2* and *P* which is what a failure of modularity would involve. Indeed, in assuming that we are learning about a single network when we do combinations of interventions, we assume the absence of such reorganization—that is, that the system is modular.

Next consider an example involving both “top-down” (as well as “bottom-up”) and “cyclic” causation. In rabbit heart cells, the potential across the cell membrane changes in a rhythmic or oscillatory fashion, with this being influenced by oscillatory behavior ion channel proteins in the cell membrane which determine the quantity of charge flowing across the membrane. The gating of the ion channels is “voltage dependent” – that is, it is causally affected by the cell potential, so that a causal feedback cycle is present between the membrane potential and the channels. Denis Noble (2012) describes this as case in which a “higher level” structure affects, via “downward causation”, lower level structures which in turn affect the higher level structure. Noble created a computer model of this interaction and then performed a “virtual” lesion or intervention on the model in which, as he puts it

The “downward causation” between the global cell property, the membrane potential and the voltage-dependent gating of the ion channels was interrupted. If there were a sub-cellular ‘program’ forcing the proteins to oscillate, the oscillations would continue. In fact, however, all oscillations cease and the activity of each protein relaxes to a steady value, as also happens experimentally. (p. 58)

In other words, Noble “broke” (his word) the downward causal link from the membrane potential to the channel gating in his computer model. Under this “intervention”, the causal link from the channels to the potential remains, but the membrane potential no longer influences (makes a difference for) the gating of the channels and their oscillatory behavior disappears. This shows that the oscillatory behavior of proteins in the intact system is indeed due to the membrane potential. Notice that (a) this is a perfectly coherent description of a system involving both bottom-up and downward causation that contains a cycle (contrary to what some have claimed about the impossibility of making sense of these notions), that (b) this pattern of dependence can be straightforwardly understood within an interventionist framework (e.g., there is downward causation because intervening on the membrane potential changes the behavior of the channels, that (c) Noble’s model is mechanistic in the sense understood in this essay and that (d) the system is understood as modular—disrupting the membrane🡪 channel gating relationship leaves the gating🡪 membrane link intact. Indeed, the modularity of the system is essential to Noble’s inference about the behavior of the intact system[[18]](#footnote-18). So contrary to what some have claimed, the modularity of a representation is consistent with that representation containing cycles or the representation of causation as operating “downwards” from upper level to lower level structures.

References

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

Bogen, J. (2004), "Analysing Causality: The Opposite of Counterfactual is Factual," *International Studies in the Philosophy of Science* 18: 3-26.

Craver, C. (2007) *Explaining the Brain*. Oxford: Oxford University Press.

Darden, L. (2006) *Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution*. Cambridge: Cambridge University Press.

Davidson, E. et al. (2002) “A Genomic Regulatory Network for Development”*Science*: 295 :1669-1678.

Dowe, P. (2000) *Physical Causation*. New York: Cambridge University Press.

Glennan, S. (2002) “Rethinking Mechanistic Explanation” *Philosophy of Science* 69: S342-53.

Greenspan, R. (2001) “The Flexible Genome” *Nature Reviews Genetics* 2: 383-7.

Hausman, D. and Woodward, J. (1999) “Independence, Invariance, and the Causal Markov Condition” *British Journal for the Philosophy of Science* 50:521-83.

Lewis, D. (1986) “Causation” *Philosophical Papers*, Volume II. Oxford: Oxford University Press, pp 159-72.

Machamer, P., Darden, L. and Craver, C. (2000) “Thinking About Mechanisms” *Philosophy of Science* 57: 1-25.

Mackie, J. (1974) *The Cement of the Universe*. Oxford: Oxford University Press

Mitchell, S. (2009) *Unsimple Truths*. Chicago: University of Chicago Press.

Noble, D. et al (1992) “Reciprocal Role of the Inward Currents ib, Na and if in Controlling and Stabilizing Pacemaker Frequency of Rabbit Sino-Atrial Node Cells”. *Proceedings of the Royal Society Lond*on B 250, 199–207.

Noble, D. (2012) “A Theory of Biological Relativity: No Privileged Level of Causation” *Interface Focus*  2, 55–64

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

Schaffer, J. (2000) “Trumping Preemption**”** *The Journal of Philosophy*, Vol. 97: 165-181.

Von Dassow, G. and Munro, E. (1999) “Modularity in Animal Development and Evolution: Elements of a Conceptual Framework for EvoDevo” *Journal of. Experimental. Zoology* 285:307-25.

Waskan, J. (2011) “Mechanistic Explanation at the Limit” S*ynthese* 183 (3):389-408.

Weber, M. (2008) “Causes without Mechanisms: Experimental Regularities, Physical Laws, and Neuroscientific Explanation” *Philosophy of Science* 75 (5):995-1007.

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

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

Woodward, J. (2006) “Sensitive and Insensitive Causation” *The Philosophical Review*.115: 1-50.

Woodward, J. (2013) “Mechanistic Explanation: Its Scope and Limits”  *Proceedings of the Aristotelian Society Supplementary Volume* lxxxvii: 39-65.

Woodward, J. (2015) “Interventionism and Causal Exclusion” *Philosophy and Phenomenological Research* 91: 303- 347.

Woodward, J. (2018) “Explanation in Neurobiology: An Interventionist Perspective” In *Explanation and Integration in Mind and Brain Science*. Ed. David Kaplan). Oxford University Press.

1. But see Woodward, 2013, for some discussion of this issue. [↑](#footnote-ref-1)
2. Of course I don’t mean that this process violates energy conservation but rather that merely observing that energy or some other conserved quantity is “transmitted” in the process tells us little about how the process works or how the chemical signal depends on the electrical one. [↑](#footnote-ref-2)
3. I will use lower case letters to refer to token events and upper case letters to refer to types. [↑](#footnote-ref-3)
4. This idea is defended in more detail in Woodward, 2006. [↑](#footnote-ref-4)
5. A number of other writers, including Craver, 2007, Glennan, 2002 and Weber, 2008 have made use of interventionist ideas to provide a characterization of mechanisms along broadly the same lines as the account that follows. I am indebted to all of them and especially to Craver, who provides an extremely clear and insightful development of interventionist ideas in the context of mechanistic explanation in neurobiology. To the extent that there is anything new in the remarks that follow, this has to do with my emphasis on the interrelations between DM and CP ideas and the role of modularity. [↑](#footnote-ref-5)
6. For some examples of non-mechanistic explanations, see Woodward, 2013. [↑](#footnote-ref-6)
7. Although I lack space for detailed discussion, much of the philosophical literature on mechanisms and mechanistic explanation fails to clearly distinguish components of mechanisms from variables that characterize the behavior of those components. This leads to a number of confusions—for example, while it makes straightforward sense to speak of a component as a spatio-temporal part of a larger system, in many cases it makes no sense to speak of the values of variables as “parts” of other variables or larger systems, even though such language is common in philosophical discussion. [↑](#footnote-ref-7)
8. For reasons of space, I will speak of systems and representations as being modular or not, but obviously modularity comes in degrees along many different dimensions: a generalization may be invariant under some changes in relationships occurring elsewhere in a system and not others, a representation may be such that some of the relationships it describes may be invariant and not others, and a relationship may be relatively invariant when described in a coarse-grained way but not more finely. A more detailed treatment would need to pay attention to all of this. [↑](#footnote-ref-8)
9. Note again the distinction between the decomposability of a system into spatial parts and the question of whether causal relationships associated with the behavior of parts of the system are independently changeable. As I understand modularity, it has to do with the latter. [↑](#footnote-ref-9)
10. Claims that an intervention has changed the value of *X* and about what would happen to the value of Y under such interventions require that the properties corresponding to *X* and *Y* exist, that the value of *X* has been changed in the way claimed and so on. One can’t intervene on what doesn’t exist and there is no well-defined notion of intervention for variables that are mere place-holders with no further interpretation. For example, there is no well-defined notion of intervention for a variable that is specified only as “whatever it is that causes *Y*”. [↑](#footnote-ref-10)
11. For an account of how to define notions of causal proximity, causal betweenenss, and “direct causation” in DM terms, see Woodward, 2003, chapter 2. [↑](#footnote-ref-11)
12. Note that in cases in which variables *X* and *Y* can take values corresponding to the absence or non-occurrence of some event (e.g,. Billy fails to throw), there will be no connecting process between these values, but there will be connecting processes between *other* values of *X* and *Y* (e.g., when Billy throws). The possibility envisioned above is that there is no connecting process for *any* pair of instantiated values for *X* and *Y*. [↑](#footnote-ref-12)
13. Consider the case in which the gardener omits to water the plants, which is represented within a DM framework by means of a cause variable, one value of which corresponds to non-watering and the other of which corresponds to watering. For the latter value there *will* be a connecting process with the plant’s survival. So this is not a case for which there is no connecting process for all values of the causally related variables, as might be imagined for magic spells. As far as I am aware this is true of virtually all genuine cases of causation by omission involving “physical” (that is, non-mental, non-social variables) —that is, the omission corresponds to one value of the cause variable but there is an alternative value of the cause variable which is related to a value of an effect variable by something like a connecting process. [↑](#footnote-ref-13)
14. This is an empirical claim which I lack the space to adequately defend here. I acknowledge that it is, at best, true only for the most part. Some relationships mediated by continuous causal processes are (at some relevant level of description) highly unstable/non-invariant—think of a golf ball that is hit (*H*), strikes a tree limb and falls into the cup (*C*), where the causal relationship of interest is between *H* and *C.* If the wind, the angle of collision with the limb and many other circumstances had been even slightly different, the ball would not have gone into the cup yet we unhesitatingly judge that *H* causes *C*. In contrast to the case just described, additional cases that suggest that stability and spatio-temporal connectedness or proximity often go together include, e.g., those in which a causal relationship is established by a physically rigid structure or by the trajectory of an ordinary cohesive physical object with non-negligible mass. Think of the very stable connection between movements imparted to the handle of a rake and the movement of the prongs or the relationship between the movement of a rigid gear in interlocking contact with another gear, which in both cases will persist under a range of conditions as long as these structures remain rigid and in contact. [↑](#footnote-ref-14)
15. I assume that explanations involve representations of a system of some appropriate kind (e.g. in terms of equations, graphs, diagrams etc.) To this extent, my notion of explanation is not an “ontic” notion. Explanations represent relationships in the world, but the explanations themselves are not “in” the worldly systems represented. [↑](#footnote-ref-15)
16. One might imagine that a complex system like a cell or a brain is modular only at some incredibly fine-grained level of resolution (e.g., individual molecules) and that, for various reasons, it is regarded as unfruitful to attempt to construct models at this level. I would count such a system as not modular in any interesting sense. [↑](#footnote-ref-16)
17. In discussing this example with others, I have encountered both reactions, endorsed with great conviction. [↑](#footnote-ref-17)
18. For a more detailed defense of the claim that one can make sense of “downward” causation within an interventionist framework, including a response to exclusion-type arguments, see Woodward, 2015. [↑](#footnote-ref-18)