In Aristotelian Society Supplementary Volume 87, Issue 1, pages 39–65, June 2013 (Symposium with John Dupre)

Mechanistic Explanation: Its Scope and Limits

James Woodward
History and Philosophy of Science
University of Pittsburgh

1.

As John Dupré's stimulating and illuminating paper (2013) attests, the notions of "mechanism" and "mechanistic explanation" are currently "hot" topics within philosophy of science. A number of writers (hereafter *mechanists*) have proposed characterizations of these notions, often with added suggestion that at least in the biomedical sciences, explanations are, in actual practice, and, as an ideal should be, mechanistic in nature or should proceed by representing biological systems as mechanisms. For these writers, biological theorizing that is not appropriately grounded in or related to mechanism information is to that extent unexplanatory. Dupré is skeptical of (at least) this last claim, writing that "there are good reasons to think that biological systems—organism, cells, pathways, etc.—are in many ways quite misleadingly thought of as mechanisms". He suggests instead the future of biological theorizing is likely to lie, at least so some considerable extent, with non-mechanistic forms of explanation, such as those provided by dynamical systems theory.

I am broadly sympathetic to (and will largely defend) many of Dupré's claims but I want to begin with a methodological (or meta-philosophical) concern. An obvious worry raised by arguments over whether there are non-mechanistic forms of explanation, either in biology or elsewhere, or whether all biological systems are "machine-like", is that such disputes are (or threaten to become) largely terminological or semantic— the answer depends on what is meant by "mechanism", "mechanical explanation" and so on and these are notions without clear boundaries. This worry is particularly pressing because different mechanists seem to have rather different views about how narrowly the relevant notions should be understood. Some, like Craver (2008) and Craver and Kaplan, (2011) appear to adopt relatively restricted notions of mechanism. In the case of mechanistic explanations in biology, these writers emphasize the importance of providing very specific, concrete details concerning, e.g., molecular interactions and their spatio-temporal organization. For such writers, dynamical systems theory and other more abstract forms of biological theorizing (such as those appealing to network structures of the sort described in Section 5 below) do not, in the absence of molecular or other sorts of low level detail, provide mechanistic explanations or indeed explanations of any kind at all. By contrast other writers, such as Bechtel (2011) adopt a less restrictive conception of mechanistic explanation and think of at least some dynamical systems-type explanations as "mechanistic" (or as providing information that contributes to the construction of mechanistic explanations). I suspect that such non-restrictionists will respond to many of Dupré's examples of putative non- mechanistic explanations by insisting that such explanations are mechanistic after all, on some appropriate (and suitably rich) conception of this notion. Restrictionists like Craver, by contrast, might well agree that Dupré's examples fail to meet suitable standards for mechanistic explanation but will regard these examples as for that reason not explanations at all, rather than as non-mechanistic explanations.

Given the desirability of avoiding this sort of terminological squabble, I propose to proceed as follows. Because I think that Dupré is correct in supposing that there are important differences between the kinds of explanations he labels mechanical and those he calls non-mechanical (differences which I also think are taken to be important by many biologists) and because we need some way of referring to those differences, I will follow Dupré in talking in terms of a contrast between mechanical and non-mechanical forms of explanation. Or, to be more precise, I will describe some explanations as " more mechanical" than others or as representing some biological systems as more "machine-like" than others. In other words, I will use "mechanical "in a restricted sense, so that at least some explanations provided by, e.g., dynamical systems theory and by various sorts of network or topological structures (see Section 5) do not count as mechanical . I emphasize, however, that this is largely a terminological convention. What is really important are the differences (as well as the similarities) among the systems and explanations discussed below, and not the particular choice of words we use to label these differences. I thus would not object in any fundamental way if someone were to insist that all of the explanations discussed below are "mechanical" in some broad sense, with the differences among them having to do with the fact that they exploit analogies with different sorts of machines (clocks versus structures with sophisticated feedback and distributed control) or different sorts of expectations about how machines will behave.

Even given this terminological decision, several issues remain. One, already alluded to, is whether we should think of the mechanical/non-mechanical contrast as a sharp dichotomy or instead more in the nature of a graded (perhaps multi-dimensional) continuum, according to which some explanations are more "mechanical" than others (in various respects). In part because the various features that I attribute to "mechanical" explanations below obviously come in degrees, I regard this "graded" conception as more appropriate. A second issue has to do with the extent to which, so to speak, mechanical and non-mechanical forms of explanation necessarily compete with or exclude one another. (Note that even if there are important differences between mechanical and non-mechanical forms of explanation, it still might be the case that biological understanding is best advanced by combining elements of both.) For reasons of space, I will not discuss this question, but my inclination is to think that a picture according to which these two varieties of explanation can reinforce or complement one another is often correct. For example, the (relatively) non-mechanical explanations that appeal to network structure discussed in Section 5 may nonetheless build on mechanism information in the sense that the network itself is an abstract representation of facts about mechanistic relationships, even though it is generic features of the network topology and not the details of those mechanistic relationships that, so to speak, drive the explanations provided.

With these preliminaries out of the way my plan is to proceed as follows. First (Section 2), I will describe two examples of machines or of systems whose behavior is represented or explained as mechanical—one a designed artifact and the other a biological system. I will then (Sections 3-4) sketch my own account of mechanistic explanation   making use of the interventionist ideas about causation I have developed elsewhere (Woodward, 2003). This account will emphasize the way in which mechanistic explanations, at least in the biomedical sciences, integrate difference-making and spatio-temporal information, and exhibit what I will call fine-tunedness of organization. It will also emphasize the role played by modularity conditions in mechanistic explanation. Section 5 argues that, given this account, it is plausible that there are forms of explanation that are relatively non-mechanical or depart from expectations one associates with the behavior of machines.

II.

Consider the following

Rube Goldberg apparatus

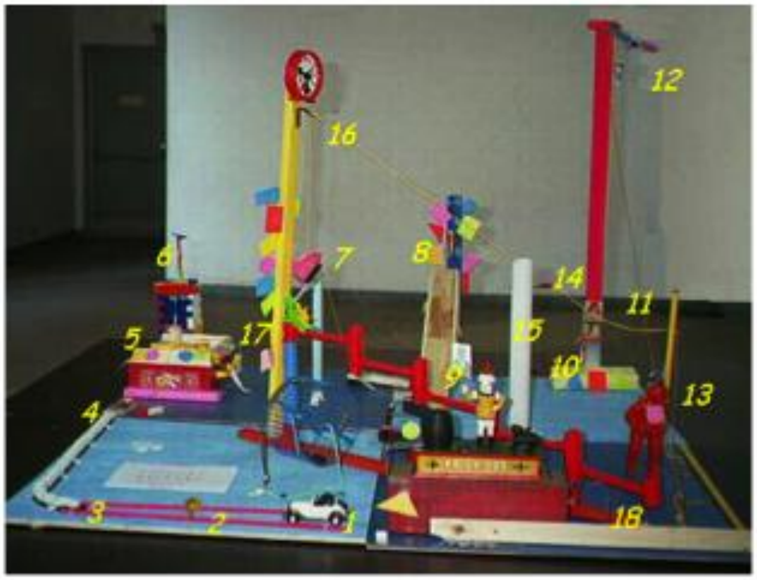


Figure 1

This structure is accompanied by the following “directions” which describe how the system is intended to work.

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!

Let us compare this with a representation of a machine-like biological structure: a diagram of the regulatory genetic network for endomesoderm specification in the sea urchin embryo, due to Davidson et al. (2002):

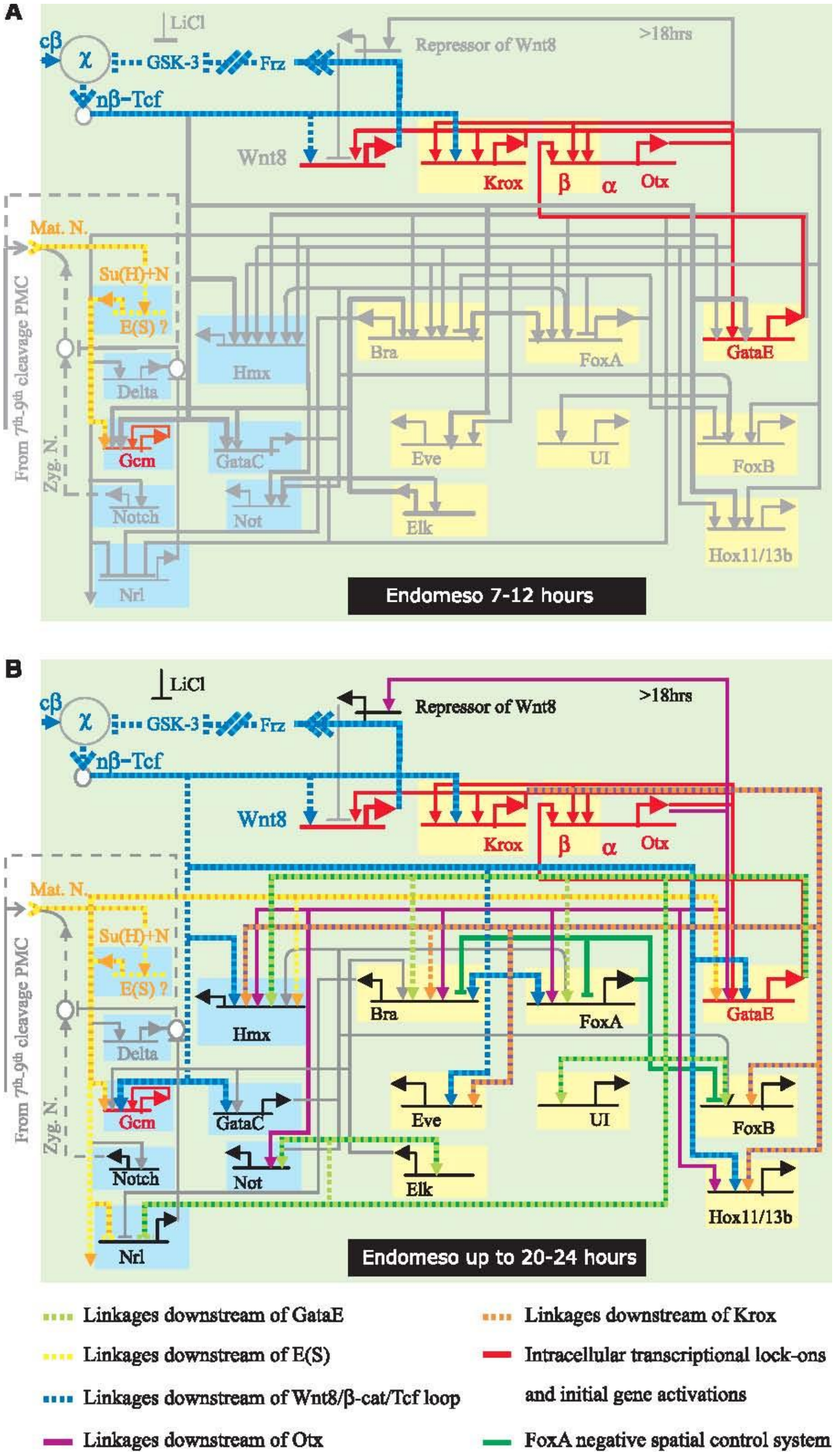

A
LiCl
cβ
χ
GSK-3
Frz
nβ-Tcf
Repressor of Wnt8
>18hrs
Wnt8
Krox
β
α
Otx
Mat. N.
Su(H)+N
E(S) ?
Delta
Hmx
Bra
FoxA
GataE
From 7th-9th cleavage PMC
Zyg. N.
Gcm
GataC
Eve
Ul
FoxB
Notch
Not
Elk
Hox11/13b
Nrl
Endomeso 7-12 hours
B
Endomeso up to 20-24 hours
Linkages downstream of GataE
Linkages downstream of E(S)
Linkages downstream of Wnt8/β-cat/Tcf loop
Linkages downstream of Otx
Linkages downstream of Krox
Intracellular transcriptional lock-ons and initial gene activations
FoxA negative spatial control system

Figure 2

Here the various arrows represent facts about relations of causal dependency—in particular that the expression of various genes or the activity of various proteins (e.g. in binding to promoter regions) influence the expression of other genes. The diagram was constructed by a "perturbation analysis" in which expression of each of the different genes or signaling pathways represented 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 whether expression (or not) of *G* made a difference for various downstream events. The arrows thus have broadly the same interpretation as the edges in directed graphs employed in causally interpreted Bayesian networks, as described in Woodward, 2003: that is, the arrows indicate that some causal dependence relation or other obtains between two relata (in the sense that there is some change in the state of one of the relata will lead to a change in the other relatum) without further specifying the details of this dependence relationship, such as its exact parameterization or functional form. The latter can be specified more precisely by means of equations -- in the example under discussion, Davidson et al. claim this can be accomplished by equations that are "Boolean" in character -- that is, by equations that involve relations like AND, OR, NOT etc. For example, such an equation might tell us that if both X and Y are expressed, Z will be expressed but not otherwise.

Davidson et al. explicitly describe this diagram as providing (or at least as contributing to) a "causal explanation" of endomesoderm development. They also add that they regard this explanation as "mechanistic" in character, writing that the diagram shows how " [I]n mechanistic terms, development proceeds as a progression of states of spatially defined regulatory gene expression".

III.

What is it about the structures in Figures 1 and 2 that leads us to think of them as representing "mechanisms" and/or as providing "mechanistic explanations? In what follows, I will assume that mechanistic explanation is a form of causal explanation and that mechanism information includes even if it is not limited to information about causal relationships[1]. Thus, as a point of departure, we need some account of cause and causal explanation. Interventionists like me (Woodward, 2003) hold that causal claims like X causes Y (where X and Y are variables) are to be understood as claims about how the value or probability distribution of values of Y would change if an intervention were to occur that changes the value of X:

(**M**): X causes Y in some background condition B_i (where background conditions are represented by variables distinct from *X* and *Y*) if and only if there is a possible intervention on *X* in B_i such that if such an intervention were to occur, the value of *Y* or the probability distribution of *Y* would change (cf. Woodward, 2003).

[1] I thus reject the idea, endorsed by some mechanists (e.g., Stuart Glennan, in his 1996) that one can appeal to the notion of mechanism to provide an account of what it is for a relationship to be causal. On my view this gets matters the wrong way around; the notion of cause is the more basic notion and information about mechanisms is information about causal relationships meeting further conditions.

Here an "intervention" is to be thought of as an idealized, unconfounded manipulation of X that occurs in such a way that Y is affected, if at all, only through this change in X and not via other causal paths that do not pass through X. Biological examples of manipulations that are (or may be) intervention-like include the gene perturbation experiments employed by Davidson et al, precisely targeted ablation experiments in neurobiology, and randomized drug trials. It seems to me that (**M**) captures in a fairly straightforward way the notion of causal dependence that is represented by the arrows in Davidson's diagram. If, say, in a perturbation that inactivates G_1 in an intervention like-way, there is a corresponding change in some downstream feature like expression of gene G_2, then we conclude that G_1 causally influences G_2 and draw an arrow from G_1 to G_2. (To anticipate a possible objection, if the effect of the perturbation of G_1 is to activate a third gene G_3 that then plays a causal role in the expression of G_2, a more complex treatment is necessary—see Section 5 below for discussion.)

As in effect noted in connection with Figure 2, the information that (**M)** is satisfied is very non-specific; it says only that Y depends in some way on X in some conditions, without telling us exactly how (in accordance with what functional form), or in precisely what circumstances Y depends on X. We can provide a more detailed characterization of the $X \rightarrow Y$ relationship by devices like equations but we can also do this by providing a more precise specification of the range of background circumstances under which the $X \rightarrow Y$ relationship is *stable* (or *invariant*)[2]. Suppose the $X \rightarrow Y$ relationship satisfies **M** for some background circumstances B_i. (In this sense, we may think of the relationship as one of intervention-supporting counterfactual dependence between X and Y). The stability of the X—> Y relationship under changes in background conditions has to do with the extent to which the relationship would continue to satisfy **M** under other background circumstances different from B_i. At one extreme, the X—>Y relationship may satisfy **M** for some B_i and yet this relationship may be highly unstable in the sense that any one of a large number of changes or departures from in this background condition would disrupt the relationship between X and Y. At the other extreme, this relationship might be relatively stable, at least for many biologically "normal" changes. True laws of nature (fundamental physical relationships like those expressed by Maxwell's equations) are extremely stable in this sense, but, in my view, few or no distinctively biological generalizations are sufficiently stable or have a sufficiently wide range of applicability to be plausibly regarded as laws. Nonetheless biological generalizations can certainly differ in their degree of stability. The relationship between possession of more than forty CAG repeats in the Huntingtin gene and development of Huntington's chorea is, unfortunately, very stable, both under environmental changes and variations elsewhere in the genome; other gene→ phenotype relations are far less stable under such changes.

Other things being equal (and they may not be), in providing explanations and in causal analysis, both ordinary folk and scientists seem (as an empirical matter) to prefer to appeal to relationships of counterfactual dependence satisfying **M** that are more rather than less stable or stable under classes of changes regarded as particularly relevant to the explanatory problem at hand. In other words, explanations appealing to more stable

[2] For more on stability and invariance see Woodward, 2003, 2006.

dependency relationships tend to strike us as deeper or more satisfying. Moreover, there are obvious normative or methodological rationales for this preference: among other considerations, more stable relationships are more general and exportable to new contexts, and knowledge of such relationships is more useful for purposes of manipulation and prediction.

Explanations that appeal to causal information that conforms to (**M**) fall into the general category of "difference-making" explanations: such explanations proceed by identifying factors which are such that changes or variations in those factors make a difference for their explananda; they answer what-if-things-had-been-different-questions. This picture of explanation allows for the possibility that difference-making factors may occur at many different "levels" of generality and abstractness. In some cases, the relevant difference-making factors may be very concrete and specific—it may be that the presence of a protein with a very specific structure at a particular concentration is what makes the difference to some biologically significant outcome, in the sense that if a protein with a different structure or at a different concentration instead had been present, that outcome would not occur. In other cases, the difference-making factors may be "higher level" and less specific—perhaps some effect will be produced as long as a neural network with some overall pattern of connectivity is present, regardless of how this is implemented in detail, or an effect will follow as long as the concentrations of various proteins falls within certain broad limits. I will suggest below that is these latter cases that often strike us as comparatively "non-mechanistic".

IV.

How might these ideas about interventions, stability, and difference-making be applied to mechanistic explanation? To fix our target, consider the following remarks by the biologists von Dassow and Munro which seem to me to capture many of the features commonly assigned to mechanistic explanation:

> 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)

Suppose we are presented with some system *S* which exhibits behavior *O* in circumstances *I*, so that the overall input/output relationship $(I \rightarrow O)$ exhibits a dependency between *I* and *O* that conforms to **M**, but where little more may be known about this relationship. For example, it may be observed that headache relief depends (in the manner described by **M**) on whether or not one ingests aspirin, that turning the handle on a cash register is followed (eventually) by turnings of the pages of a book as in Figure 1, or that development in some organism follows a certain time course linking earlier to later stages. I contend that at least in many case mechanistic explanations of the behavior of *S* exhibit the following structure: First, *S* is revealed to consist of components or parts, characterized by variables $X_1, X_k,$ where these variables themselves stand in causal relationships (again in the sense captured by **M**) that serve as intermediate or intervening links along causal paths running from *I* to *O*. In some cases, the structure instantiated by these intermediate steps may be very simple. For example in figure 1, the causal structure of the apparatus is one in which in the Input variable *P* = position of handle on cash register causes *R*= movement of register drawer causes *M*= motion of golf ball in a sequence leading to the page turnings *T*: $P \rightarrow R \rightarrow M \rightarrow \ldots \rightarrow T$, where each of these intermediate causal relationships conforms to **M**. In the diagram constructed by Davidson et al. the intervening causal pathways are much more complex, but the same underlying principle of identifying mediating or intervening causal relationships among components is at work. In such more complex cases, an intervening structure in which various sorts of diverging and converging paths will be present—e.g., *I* causes X_1 which causes X_2 and X_3, X_2 and X_3 cause X_4 and so on.

Phenomenologically, this provision of information about intermediate causal links gives one the sense of opening up the black box represented by the overall $I \rightarrow O$ relationship. The original relationship may strike one as arbitrary or puzzling (or at least one may not understand why it holds): why should ingesting a small white pill make one's headache go away, why should pulling on a register handle turn the pates of the book? Revealing the intermediate links and intervening variables by which *I* causes *O* serves to help to dispel this puzzlement, providing what seems like a "deeper" explanation for why the *I*—> *O* relationship holds.

But why (or in virtue of what) should we think the information just described as furnishes a deeper understanding of what is going on? Here I would appeal to three, interrelated ideas, which I will discuss in turn: 1) Stability, 2) Modularity, and 3) Fine-Tunedness of Organization, spatio-temporal and otherwise.

4.1) *Stability.* In successful mechanistic explanations, the generalizations describing intermediate links are typically expected to be (and often are) more stable than the overall $I \rightarrow O$ relationship with which we begin. There is an obvious although defeasible reason why we should expect this. Suppose, to take a particularly simple case, there are two intervening variables X_1 and X_2, so that $I \rightarrow X_1 \rightarrow X_2 \rightarrow O$. Absent the operation of additional factors, the $I \rightarrow$O relationship will only be stable under changes that fail to disrupt *any* of these four intervening links —the chain is no more stable than its least stable link. By contrast, if, as often will be the case, some of the changes in background circumstances under which the $I \rightarrow X_1$ relationship is stable are different from the changes in background circumstances under which $X_1 \rightarrow X_2$ is stable, and similarly for the other intermediate links, then , ceteris paribus, it will be reasonable to the expect that each of these intermediate links, taken individually, will be more stable than the overall $I \rightarrow O$ relationship. For example, in the case of the Rube Goldberg machine, the overall relationship between pulling the register handle and the turning of the book pages is (one would suppose) highly unstable—anyone of a large number of changes (such as cutting of strings or wires, tilting the platform on which the golf ball rests so that the ball does not fall off its edge, shaking the apparatus vigorously etc.) will disrupt this relationship. (This instability can be connected to the explanatory unsatisfactoriness or shallowness, noted above, of appealing to the $I \rightarrow O$ relationship by itself as an explanation of anything.) By contrast, the intervening links in the operation of the apparatus, taken individually, are more stable, often considerably more so. Insofar as the exhibition of more stable dependency relationships leads to deeper explanations, this consideration helps to account for our sense that the decomposition of the behavior of the apparatus into intervening links yields greater understanding of what is going on. I take this greater stability of intervening links to be a characteristic feature of at least one kind of simple, machine-like behavior.

In the case of Davidson's diagram, the individual causal links reflect facts about biochemical reactions and the rate equations that underlie these. If one were to take Davidson's diagram at face value (that is, as a correct and complete description of the causal structure of the system), it seems natural to expect that these individual links will be more stable than the overall developmental trajectory, on the basis of the same considerations that were at work in connection with the Rube Goldberg machine. If, as I take to be a genuine empirical possibility, this is not the case (and the overall trajectory of development is instead more stable and robust than the individual links—see Section 5) this might be interpreted as an indication that Davidson's diagram has left something important out—perhaps there are other causal relationships, not represented in the diagram, which operate to sustain a robust developmental trajectory when the relations represented in the diagram are disrupted. I thus suggest that to the extent that the behavior of a system is the result of causal relationships among components that are individually less stable than the system's overall behavior one has departed from one kind of expectation that is naturally associated with simple machine like behavior.

4.2. *Modularity.* Modularity conditions can be regarded as one particular kind of stability condition, in which our concern is with whether individual causal relations are stable under changes in other sets of causal relationships[3]. Suppose that an overall $I \rightarrow O$ relation characterizing system *S* is represented as decomposed into intervening causal relationships among the components of *S*, where these relationships are described by a set of generalizations $\{G_i\}$. I will say that this representation is modular to the extent that each of the individual G_i remain at least somewhat stable under interventions that change the other G_i. Put informally, modularity (in this sense) has to do with the extent to which one can change or disrupt the causal relationships governing one subset of components in the system without changing or disrupting the causal relationships governing other, distinct subsets of components. To the extent that an explanation/representation of the behavior of a system in terms of its components is modular, we can associate distinct operations or patterns of behavior or interaction with different subsets of components—each such subset of causal related components continues to be governed by the same set of causal relationships, independently of what may be happening with components outside of that subset, so that the behavior is (in this respect) "intrinsic" to that subset of components. (Compare the quotation from von Dassow et al. above, with its emphasis on the "intrinsic" behavior of components—modularity seems to me to capture at least part of what is involved in their notion of "intrinsicness".) To the extent this is the case, we can move a modules from one causal context to another and expect it to operate "in the

[3] There are a number of (apparently) different notions of modularity that are employed in biological discussion, with independent disruptability being just one of these. For example, structures whose parts or nodes are relatively densely casually connected and which are only sparsely connected to other such structures are also often described as modules. It is an interesting question how this notion relates to the independent disruptability notion on which I focus above.

same way". The extent to which an explanation or representation is modular is, I believe, another dimension which is relevant to whether the explanation is "mechanical" or exhibits the system of interest as behaving in a machine-like way.

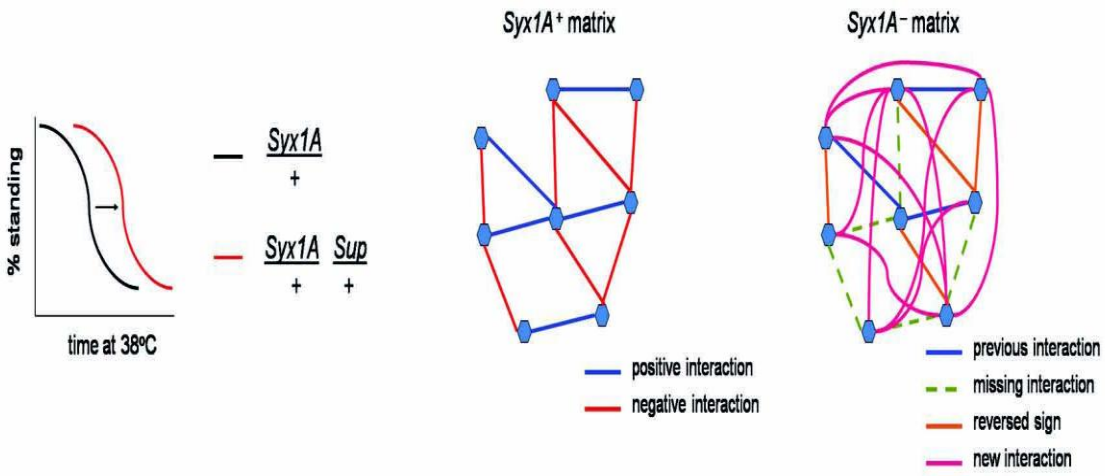
Since this notion of modularity is easily (and often) misunderstood, it is worth spelling out in a bit more detail. First, as I think of modularity, it is a feature of *representations* or *explanations* of the behavior of systems, rather than of the system themselves. This allows for the possibility that one particular representation *R* of system *S* might be non-modular and another, different, perhaps more fine-grained representation *R** of the same system *S* might be modular. (A representation of the brain in terms of coarse-grained psychological variables like "cognitive judgment" and "emotion" might be highly non-modular, a representation at the level of individual neurons more modular.) Second, like stability, modularity comes in degrees and is relative to a class of changes or interventions. Generalization G_k characterizing the causal relationships among one subset of components might be stable under certain sorts of changes C_i in the causal relationships characterizing other subsets of components but not under other changes $C_j \neq C_i$ in these relationships. Or G_k might be stable under changes in some causal relationships elsewhere in the system but not under changes in other causal relationships. Or perhaps changes elsewhere change G_k but only in small or minor ways.

Third, it is important to understand that modularity does *not* imply that if we disrupt some upstream component in a mechanism, the downstream components will continue to produce the same outputs as before. Rather modularity has to do with the extent to which causal *relationships* governing subsets of components remain stable under changes in the relationships governing other subsets of components. Thus if the representation in figure 1 is modular, this implies that it is possible to disrupt or change the relationship described in step 3, perhaps by cutting the string that releases the bus in the intact mechanism, without altering the causal relationships that characterize the operation of other parts of the mechanism—for example, the relationship between the movement of the bus and the movement of the ball described in step 5. More specifically, if one were to cut the string in step 3, but then intervene to move the bus down the ramp as described in steps 4-5 so that it strikes the ball appropriately, then (to the extent that modularity is satisfied) the rest of the apparatus would continue to function as before—that is, the causal relationship in step 5 would continue to hold. Of course if the string in step 3 is cut and no other changes occur, the inputs to all steps downstream from step 3 will be altered and the pages of the book will not be turned, but modularity is consistent with this. Similarly, if the representation in figure 2 is modular, this implies that it is possible to alter the relationship between, eg, gene 1 and gene 2, without altering other relationships between other genes elsewhere in the system. Suppose that G_1 influences the level of expression of G_2 in accord with some generalization g_{12} and that G_2 in turn influences G_3 in accord with generalization g_{23}. A mutation in G_1 to G_1 * leads to a new generalization g_{12}* describing the relationship between G_1 * and G_2. Such a mutation may change the level of expression of G_2, but this is consistent with the system being modular. What modularity requires is that it be possible for the change from g_{12} to g_{12}* to occur without a change in generalization g_{23}. In terms of a graphical representation, this amounts to the idea that one can disrupt the arrows directed into one variable or node without disrupting or changing arrows directed into another variable.

It is also important to understand that the claim that an explanation/representation is modular does *not* require that causal interactions be linear or additive. Suppose that X_1 and X_2 cause X_3 through some non-linear interaction characterized by g_{123} $(X_1, X_2) = X_3$. As long as changes in g_{123} leave relationships elsewhere in the system involving other variables intact and vice-versa, the system counts as modular. As an illustration, consider a hypothetical mechanism, which contains, as a component, a match which is struck or not (*S*) in the presence or absence of oxygen (*O*), leading to ignition or not (*I*). When the match ignites, it is used to start a fire (*F*) which in turn boils water (*B*). The overall system is described by a set of generalizations, g_1 which describes the relationship between *S*, *O* and *I*, g_2 which describes the relationship between *I* and *F*, and g_3 which describes the relationship between *F* and *B*. The relationship g_1 is non-linear or non – additive and involves an interaction effect: Striking the match has different effects on *F*, depending on whether oxygen is present. Nonetheless this representation may still be modular: even if the relationship between *S, O* and *F* is disrupted (e.g. by soaking the match in water), the relationship between *I* and *F* and between *F* and *B* may remain intact—if we could get the match to ignite in some other way it would still start a fire which would boil the water.

I emphasize this point because there is little doubt that many of the causal relationships that are present in biological systems are highly non linear and non-additive. If modularity implied linearity or additivity or the absence of interaction effects among components, it would not be a very useful notion for characterizing biological explanations. And if all mechanical explanations must satisfy a modularity condition and modularity requires additivity, we could conclude immediately that there are very few mechanical explanations in biology. This anti-mechanist conclusion strikes me as far too quick; my contrary view is that there is no reason to hold that explanations involving non-linear or non –additive relationships are for that reason alone non-mechanistic.

What then would constitute a failure of modularity? Consider the following hypothetical genetic regulatory network which is taken from Greenspan 2001 and discussed in Mitchell, 2009.



The diagram on the left represents the form originally taken by the network. Greenspan supposes that an experimental intervention in the form of a knock-out experiment on the gene/node in the upper left guardant results in a global re-organization of the whole network, so that links elsewhere in the network, not involving the gene intervened on, are altered or removed and new links are introduced. To the extent that this is the case, the representation of the original network is to that extent non- modular. This failure of modularity represents a violation of at least some of the expectations one naturally associates with mechanistic explanation. Greenspan's diagrams do not conform to the ideal of explaining how a system behaves by decomposition into components and behaviors intrinsic to these components. Instead, at least as represented, how some of the components behave depends in a global, "extrinsic" way on what is going on in other components.

I take it to be an empirical question whether the behavior of some biological system is best explained by means of a theory or representational structure that is relatively modular. It seems to me plausible that for some systems the answer to this question is "no"—see section 5 below. However, in my view it would be a mistake to dismiss the notion of modularity or to regard it as biologically uninteresting merely because it fails to have universal application. Assumptions about modularity of various sorts are pervasive in contemporary biology, across contexts as different as genetic regulation, networks of protein interaction, and neurobiology[4]. Davidson et al. in the paper referred to above, are quite happy to think of genetic regulatory systems as composed of "modules" Moreover, they note that the project of " inferring "normal function from the results of perturbation" requires what are in effect modularity assumptions of various sorts—one needs to assume that when one intervenes locally on one set of components, other components continue to operate in the same way if one is to learn anything about the intact system from such interventions As the contrast between systems in which such assumptions are satisfied and the system in Greenspan's diagram 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.

4.3. *Organization Sensitivity/Fine -Tunedness*[5]. I turn now to another feature that is present in at least many machines and mechanical explanations, which I will call

[4] To take a simple illustration, the fact that structures involved in gene regulation are often modular in the sense that they are capable of changing (and do change) independently of the structural genes that they regulate has important implications for understanding evolution.

[5] In a forthcoming paper that I became aware of as I was working on this essay, Levy and Bechtel also emphasize the contrast between, on the one hand, (1) explanations that identify very specific factors having to do with the detailed causal properties of components and the details of their spatio-temporal connections and, on the other, (2) explanations that instead abstract away from those details, such as those that appeal to general patterns of network connectivity as difference-making factors. If I have understood them correctly, they regard both sorts of explanations as in some broad sense mechanistic, but putting this difference with the position taken in this essay aside (and, as noted above, the difference may be largely terminological), their attention to the difference between (1) and (2) and their view that explanations can take the form of (2) and not just (1) is very similar to the position taken in this essay. Levy and Bechtel also make the important point that explanations that exhibit fine-tunedness of kind (1) often

organization sensitivity or fine-tunedness. Roughly speaking, this is the idea that it is characteristic of many machines that the specific details of the causal properties possessed by the components, and the way in which those components are connected to each other, including their spatio-temporal relations, matter a great deal to the outcome produced, in the sense that if these features or organization were different, the outcome would be different. For example, if one were to take the components that make up the Rube Goldberg apparatus in Figure 1 and significantly change their spatial relations, we would expect to change the input/ output relationship characterizing the original apparatus in a very substantial way. A similar result would be expected, if one were to replace one or more of the components with other, different components with different causal properties. A similar point holds for changes in the causal characteristics of the components in the network described in figure 2. In such a network, it is assumed that only very specific proteins will bind to promoter regions of particular genes and affect their regulations; replacing these with other proteins will greatly affect gene regulation. Similarly, physical interaction/contact of protein and promoter region is required; if the spatial layout of the regulatory network is altered in such a way that the appropriate protein is not able to reach the promoter region, this will alter gene expression and subsequent developmental events.

Not only is the overall behavior of typical mechanisms sensitive to the details of their spatio-temporal organization, but we also expect that there will be systematic relationships or correspondences between the causal relationships that characterize such devices (when these are understood, as above, in interventionist terms) and spatio-temporal facts about the relationships among components. To take a very simple possibility if there is a causal chain running from $X_1 \rightarrow X_2 \rightarrow X_3$, we naturally expect X_2 to occur or be instantiated temporally between X_1 and X_3, and when direct causal interaction requires spatial contact, we expect to find the component realizing X_2 spatially between the components realizing X_1 and X_3. Failure to find such expected components or instantiations at appropriate spatio-temporal locations will often cast doubt on a proposed mechanistic explanation.

Another connection between difference-making and spatio-temporal organization concerns modularity. Typically, in mechanical or machine-like devices the independent disruptability of causal relationships that is characteristic of modular organization is paralleled by the spatial (or temporal) separation of the instantiation of those relationships. For example, our confidence that the device represented in Figure 1 is modular is connected to and supported by the observation that the different components are spatially separated in a way that allows for, e.g., interference with the relationship between the register handle and the movement of the golf ball without interference with the relationship between the movement of the bus and the rubber ball in step 4. Of course spatial separation does not guarantee modularity but modularity is typically less likely to be present in the absence of the right sort of spatial and/or temporal separation. Put differently, facts about spatio-temporal separation can give us

also describe interactions among components that have very different causal properties rather than interactions among components all of which have broadly the same properties such as molecules in a gas.

clues about how to decompose a system into parts and causal relationships governing these that are independently changeable, thus facilitating "mechanistic" explanation.

This sensitivity of typical machines to details of organization – to spatio-temporal relationships among components, and to just which components are present—is of course connected to and motivates the idea that it is important to make such details explicit in constructing mechanical explanations. For example, "mechanistic" details about just which enzyme is involved in the synthesis of a particular protein and about exactly where such synthesis occurs are of scientific interest at least in part because of the assumption that, were a different enzyme involved or were it delivered to a different location, the behavior of the system would be different. To the extent that such details do not matter—to the extent that there is reason to think that some system would behave in the same way (for some behavior of interest) even if the spatio-temporal organization of the components were changed considerably or even if some or many of the components were replaced with others with different causal properties—then the motivation for constructing explanations that focus on them is correspondingly diminished, at least on a conception of explanation according to which this involves the exhibition of difference-making factors[6]. Again, it seems to me that to the extent the behavior of a system is not such that it exhibits sensitivity to details of organization of the sort I have been describing, then it departs from at least some of the expectations we associate with machine-like behavior.

V.

So far we have identified several features that are characteristic of at least many mechanistic explanations, including decomposition into relatively stable sets of intermediate causal relationships involving spatio-temporally distinguishable components, where these intermediate causal relationships satisfy modularity conditions, and where specific details of the organization and causal properties of components matter to the outcome produced. I will assume in what follows that to the extent that a purported explanation of the behavior of a system does not exhibit these features, that explanation is not mechanistic, or at least comparatively less mechanistic than explanations that do have these features.

In thinking about such non-mechanistic possibilities, an important point of departure is the observation that may biological systems, from genetic regulatory networks to metabolic networks to nervous systems, exhibit a high degree of robustness which one can understand generically as preservation of function or some overall behavior under both internal changes and external environmental changes (cf. Kitano, 2004). I believe that it is in thinking about the various ways that such robustness of behavior might be explained that the possibility of explanations that lack some of the features often expected of mechanical explanations comes to the fore.

[6] In other words, we should distinguish the idea that we should cite such lower level details when they make a difference but not necessarily when they do not, from the idea that explanations that cite such details are always better, regardless of whether they make a difference, because they are more "fundamental" or because the real site of causal action is always at the level of basic chemistry and physics.

At a very general level one may distinguish several different ways in which biological robustness can be achieved. One possibility is that some behavior *B* is ordinarily produced by structure S_1 but that when S_1 is disrupted, another, distinct back-up or fail safe system S_2 becomes operative which then produces *B*. This is how many machines of human design, such as aircraft control systems, achieve robustness. In the biological realm, what is sometimes called genetic redundancy has a similar character[7]. A surprisingly large number of genetic knockout experiments produce no noticeable phenotypic effect, even in cases in which it is known that the inactivated gene G_1 causally contributes to some phenotypic trait *P* in normal functioning. In cases involving straightforward redundancy, this is due to the fact that some back up gene G_2– perhaps a duplicate or near duplicate of G_1—becomes operative when G_1 is inactivated and contributes to *P*.

Cases in which robustness is achieved through redundant back-up mechanisms conform fairly well to the requirements on mechanistic explanation described above. As I explain in more detail elsewhere (Woodward, forthcoming), such examples can be given a straightforward interventionist treatment in terms of the behavior of the system under combinations of interventions and this maps into what might be done experimentally to disentangle the causal structure of a system with redundancy: for example, one may detect the back-up role of G_2 by intervening to knock out G_1 and simultaneously "wiggling" G_2 and observing the consequences for *P*. In this sort of case, one still has the characteristic mechanistic pattern of decomposition of an overall input out relation into intermediate causal links organized, where the details of how these are organized make a difference to the outcome achieved. Moreover, structures of this sort respect modularity: In carrying out the interventions described above, we assume that knocking out G_1 does not change the relationship between G_1 and G_2 or between G_2 and *P* which is what a failure of modularity would involve. Typically in such cases this modular organization is paralleled by a spatial separation of G_1 and G_2 — this helps to makes it plausible that the link between G_1 and *P* can be disrupted without disrupting the $G_2 \rightarrow P$ link and conversely.

I think that the most plausible examples of behavior that is not best explained mechanistically are cases of robust behavior for which the explanation of this behavior does not have to do with redundancy or the operation of well-defined back-up systems, but rather has some other basis. Here are some possible cases:

5.1) *Robustness without Fine-tunedness.* A system may exhibit robust behavior which is not the result of the fine-tuned organization of parts and components, and which

[7] Wagner (2005) clearly distinguishes between genetic redundancy in this sense and what he calls "distributed robustness" as alternative ways of achieving biological robustness. The latter involves systematic or network reorganization and/or structures (as when blocking a certain part of metabolic network leads to increased flow through other parts of the network in compensation) and/or structures whose output is not sensitive to exact values taken by variables, parameters or interactions (as when eliminating chemical reactions from metabolic networks has little effect on output or when a genetic regulatory network operates in the same way even though biochemical parameters characterizing the network change by an order of magnitude (p. 177). The latter are good illustrations of what I call robustness without fine-tunedness below.

is also not the result of the operation of some causally and spatially distinct back-up module, as in cases of genetic redundancy. Instead, the robust behavior is the result of the fact that the system is such that as long as many of the variables characterizing the system fall within certain very broad ranges the system exhibits the same behavior and similarly for variations in spatial organization. Spindle formation in mitosis, as portrayed by Dupré, is apparently an illustration of this. It turns out that there are many different pathways or causal factors involved in spindle formation; when one of these is disrupted, others continue to operate in such a way that normal spindle formation is achieved. Moreover, unlike cases of genuine redundancy, this is not a matter of one or more causally and spatially distinct back-up structures becoming activated when a primary structure is inactivated; instead the different factors influencing spindle formation are continually interactive and co-mingled, mutually modifying one another. On Dupré's view, the achievement of spindle formation is instead more like (or more appropriately modeled as) the settling of high dimensional system with a complex dynamics subject to many different constraints into a basin of attraction for the system. It is characteristic of this sort of dynamical explanation that it shows us that the dynamics of the system are such that the system will settle into the same outcome state for any one of a very large set of combination of values for the state- variables characterizing the system—in this sense the particular values taken by those variables or the particular initial conditions in which the system begins or its particular trajectory through phase space which leads to the final outcome do not matter, at least to the extent that all we want to explain is why that outcome is achieved, one way or another. To the extent that it is characteristic of mechanical explanation that it involves an exhibition of specific details of organization or particular values of variables and parameters that do "make a difference" to the behavior of the system, such dynamical-systems-type explanations do not seem very mechanical. Such dynamical systems explanations also don't seem to proceed by the decomposition into parts with intrinsic behaviors that is characteristic of modular styles of explanation[8].

As a second example, consider, (following Alon 2007) two competing models for the explanations of "exact adaptation" in bacterial chemotaxis. Adaptation refers to the ability of bacteria to adjust their motion in the presence of attractants (or repellents) in such a way that their movements become independent of attractant levels. One possible explanation, involving what Alon calls a "fine-tuned" model, takes adaptation to result from "a precise balance of different biochemical parameters" (p. 143). In this model, adaptation depends, among other things, on the precise level of the enzyme CheR which influences receptor activity (and hence motion) through methylation. A modest change in the level of CheR undercuts adaptation The second model, due to Barkai and Leibler

[8] Another striking example of a dynamical systems explanation can be found in Huang et al. (2005). In this paper, distinct cell fates are modeled as attractor states in a very high (2773) dimension gene expression state space. The authors showed experimentally that triggering divergent patterns of gene expression resulted in initially divergent cell differentiation followed by convergence to a common outcome across many different patterns of gene expression. Again, the apparent message is that the details of the particular trajectory followed do not matter as long as the trajectory is in the basin of attraction of the dynamics. In this sense the explanation for cell fate in the paper is non-mechanical.

(1997), instead takes adaptation to be “robust” and “ to occur for a wide range of parameters”, including varying levels of CheR. (Alon, 143). In this model, methylation by CheR occurs at a constant rate and adaptation occurs “because the concentration of active receptors adjusts itself so that the demthylation rate is equal to the constant methylation rate” (146). If one expects that the explanation of adaptation in chemotaxis will involve structures that are like those present in machines like clocks or the device in Figure 1, one presumably will be inclined toward the fine-tuned model. Nonetheless, experimental evidence supports the second, “robust” model. Again, the second model departs from the “mechanical” paradigm of fine-tunedness, although in other respects it may still seem “mechanical” – for example, it describes how overall behavior results from components (receptors, enzymes etc.) and causal interactions among these, where these interactions are characterized by locally stable generalizations (e.g. equations specifying how the concentration of active receptors depends on enzyme concentrations.)[9]

Topological or network explanations. A second set of explanations that seem in at least some respects non-mechanical by the standards described above appeal to what might be described as topological considerations, where this notion is used in an extended way to include explanations that appeal to patterns of connectivity, as in graphical or network structures. (I take the notion of topological explanation from the very interesting discussion in Huneman, 2010. I will add that I do not mean to suggest that such explanations will sharply distinct from those discussed under the Non-Fine – Tunedness heading immediately above). The characteristic idea behind such explanations is that the difference-making factors for certain explananda are relatively abstract systems level facts about, e.g., generic features of patterns of connectivity in which components stand or about the overall shape of the space of solutions of the equations governing the system, rather than either more specific causal properties of those components or very specific details of spatio-temporal relationships among those components. Thus it may be that as long as the components interact (in any way) that conforms to a pattern of interaction that exhibits a certain topology , various outcomes follow. Or it may be that the difference-making features involve the combination of the pattern of overall connectivity and certain qualitative features of the connections such as whether they are excitatory or inhibitory, but not the precise functions governing the behavior of the components. In ecology, networks of interaction among species may be specified independently not just of the specific species standing in those network relations, or the specific causal details of how these species interact, but even independently of whether the occupants of these roles act as predator or as prey. It is then suggested that as long as such networks have various generic properties they will have various other properties of interest to ecologists. For example, networks that are scale –free (that is, that exhibit the same pattern of connectivity at all scales) are conjectured to have the property of being relatively robust against random deletions of nodes and connections (corresponding to extinctions in ecological contexts) . Similar explanations that appeal to generic features of networks have also been used in neurobiology (e.g., Sporns, 2011), in a project of classifying different possible sorts of neural networks in terms of generic connectivity properties and then investigating other features that are present just as a consequence of

[9] In fact Alon himself describes the second model as the Baikai-Leibler “mechanism”.

these connectivity properties. For example, many neural structures seem to be organized as "small world" networks in which most nodes are not directly linked to one another but nonetheless can be reached by relatively short paths. Such networks are conjectured to have, merely in consequence of their small world organization, a number of other features of neurobiological interest—for example, in addition to being unaffected by random deletion of edges, as is also the case with the ecological networks mentioned above, such small world networks are also thought to be more evolvable than alternative forms of network organization. As is perhaps obvious, accounts that appeal to network or topological properties are particularly appealing insofar as what we are trying to explain are why biological systems exhibit the robustness properties that they do, since the characteristic feature of such explanations is that they show us that as long as the relevant network/topological features are preserved, other sorts of changes or perturbations to the systems in question do not affect their behavior[10].

An obvious question that arises in connection with such examples, which is particularly likely to be pressed by those who favor restricted conceptions of mechanistic explanation, is whether we should we think of them as explanations at all. This is large question that deserves a more detailed answer than I can provide here, but, very briefly, I would argue for an affirmative answer by appealing to the general interventionist framework for understanding explanation described in section 3. In particular, successful dynamical systems and topological explanations have the following properties: First, they locate difference-making features for (or answer what-if-things-had- been different questions with respect to) their explananda, exhibiting dependency relationships between those explananda and factors cited in their explanantia. For example, an explanation that appeals to network structure may show us that changing generic features of the connectivity of a network will change the outcomes produced by the network, while changes within the network that do not change this connectivity will have little or no effect. Second, such dependency relationships are, in the examples discussed, stable or invariant, in the sense that the relationships continue to hold under ranges of changes in other background conditions—e.g., as long as an network ecological network is scale-free, it may be robust under deletion of edges, independently of a range of other changes. Sometimes these dependency or difference-making relationships can be given an interventionist interpretation in the sense that they encode information about what would happen under possible manipulations. For example, one may think of many explanations that appeal to network structure as telling us what would happen if one were to manipulate generic features of that structure, by removing or adding edges. In other cases, although one can think of the dependency relations in questions as telling us what would happen if various factors (e.g., having to do with network structure or the shape of an attractor landscape) were different, but the notion of interventions or manipulations that change those factors may make less sense—perhaps such manipulations are either not physically possible or because they are unlikely to meet the conditions for a local, surgical intervention. In such cases, one may have difference-making relations or dependencies without an interventionist interpretation, as briefly discussed in Woodward, 2003 pp. 220-1.

[10] This point is emphasized by Huneman, 2010.

Throughout this paper I have resisted the impulse to enlarge the notion of mechanism in such a way that all explanations in biology automatically turn out to be “mechanistic”. Only by adopting a notion of mechanism that is restricted in some way can one even raise questions about the scope and limits of mechanistic explanation. My view is that mechanistic explanations are most likely to be successful when the systems to which they are applied satisfy certain empirical presuppositions, having to do with the possibility of decomposition into stable intervening links, modularity, fine-tunedness and so on; as one moves away from contexts in which these presuppositions are satisfied, mechanistic explanation becomes a less promising strategy. This is not intended as a criticism of mechanistic explanation in those circumstances in which it is appropriate, but rather simply reflects the point that, like virtually all explanatory strategies and structures, there are limits on its range of application.

References

Alon, U. 2007: *An Introduction to Systems Biology*. London: Chapman and Hall.

Barkai, N. and Leibler, S. 1997: “Robustness in Simple Biochemical Networks” *Nature* 387: 913-917.

Bechtel, W. 2011: “Mechanism and Biological Explanation” *Philosophy of Science*, 78: 533–557.

Craver, C. 2008: “Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential” *Philosophy of Science*. 75: 1022–1033.

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

Dupré, J. 2013: “Living Causes” *Aristotelian Society Supplementary volume*

Glennan, S. 1996: “Mechanisms and the Nature of Causation” *Erkenntnis* 44, 49-71

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

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. 2005: “Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network” *Physical Review Letters* 94: 128701.

Huneman, P. 2010: “Topological Explanations and Robustness in biological sciences” *Synthese* 177: 213-245

Kaplan, D.M. and Craver, C.F. 2011: “The Explanatory Force of Dynamical Models” *Philosophy of Science* 78: 601-627.

Kitano, H. 2004: “Biological Robustness” *Nature Reviews Genetics* 5:826-37.

Levy, A. and Bechtel, W. Forthcoming: "Abstraction and the Organization of Mechanisms" *Philosophy of Science***.**

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

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

Sporns, O. 2011: *Networks of the Brain.* Cambridge, MA: MIT Press.

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.

Wagner, A. 2005: "Distributed Robustness versus Redundancy as Causes of Mutational Robustness" *BioEssays* 27: 176-188.

Woodward, J. Forthcoming:"Causation and Mechanisms in Biology" To appear in a volume of *Minnesota Studies in Philosophy of Science*.

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

Woodward, J. 2006: "Sensitive and Insensitive Causation" *Philosophical Review* 115: 1-50.