What Is a Mechanism? A Counterfactual Account

Jim Woodward†‡
California Institute of Technology

This paper presents a counterfactual account of what a mechanism is. Mechanisms consist of parts, the behavior of which conforms to generalizations that are invariant under interventions, and which are modular in the sense that it is possible in principle to change the behavior of one part independently of the others. Each of these features can be captured by the truth of certain counterfactuals.

1. Introduction. My co-symposiasts, Lindley Darden and Stuart Glennan, argue that the construction of explanations and the identification of causal relationships is closely bound up with the discovery of mechanisms. I fully agree. In what follows I will try to show how ideas that I have recently defended linking causal and explanatory relationships to relationships that are invariant under interventions can also be used to provide a characterization of the notion of a mechanism.

What is a mechanism? In a recent paper, Machamer, Darden, and Craver (hereafter MDC) write:

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (2000, 3)

Elsewhere they write that descriptions of mechanisms are explanatory in virtue of revealing “productive relations,” and that rendering phenomena within a given area “intelligible” has to do with “portraying mechanisms

†Send requests for reprints to the author, Humanities & Social Sciences, H&SS, 101–40, California Institute of Technology, Pasadena, CA 91125; e-mail: jfw@hss.caltech.edu.

‡Thanks to my co-symposiasts Lindley Darden and Stuart Glennan for helpful comments.

Philosophy of Science, 69 (September 2002) pp. S366–S377. 0031-8248/2002/69supp-0031$10.00

in terms of a field's bottom out [that is, fundamental] entities and activities" (2000, 3). For example, bottom out activities in molecular biology and molecular neurobiology fall into the following four categories: geometrical-mechanical, electro-chemical, energetic, and electro-magnetic.

While this is unexceptionable as far as it goes, I think that it is worth exploring whether it is possible to provide a more general and less discipline specific characterization of notions like "mechanism" and "production."

2. A Physical Example. It will be useful to begin with a concrete example. Consider (Ex1) a block sliding down an inclined plane. The standard textbook analysis proceeds as follows. The block is subject to two forces—a gravitational force due to the weight of the block and a force due to friction which opposes the motion of the block. The frictional force F_k obeys the relationship

$$F_k = \mu_k N \tag{1}$$

where μ_k is the coefficient of kinetic friction and N is the normal force exerted perpendicular to the direction of motion of the block. From the above diagram, we have that the component of the gravitational force due to the weight of the block along the plane is mg sin ø. The normal force $N = mg \cos ø$ and so the frictional force $F_k = \mu_k\, mg \cos ø$. The net force on the block along the plane is thus

$$F_{net} = mg \sin ø - mg \cos ø \tag{2}$$

and the acceleration a of the block is given by

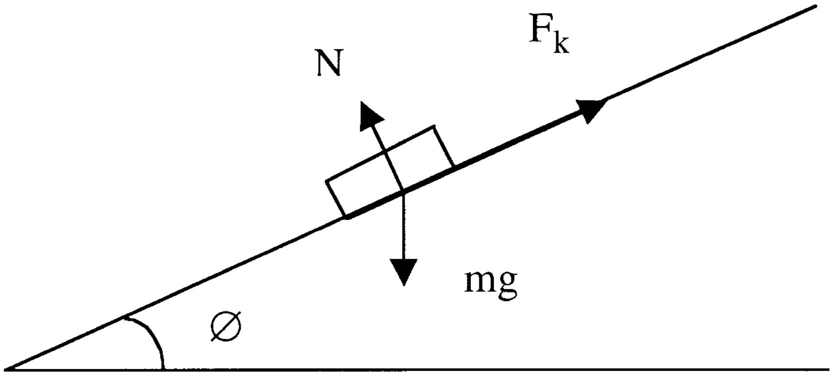

Figure 1

$$a = g \sin ø - g \cos ø. \tag{3}$$

Although this is a particularly simple example, it exhibits a number of features possessed by more interesting and complex mechanisms. The first point to notice is that like other mechanisms, this one is made up of parts or components that, as MDC say, are "productive of regular changes." What does this mean? It is a familiar point that not every generalization describing a regularity describes a productive or causal relationship. Even though there is a correlation or regular association between the joint effects of a common cause, neither of these effects produce regular changes in the other—a point that is illustrated by the familiar example of the correlation between the reading of a barometer B and the occurrence/non-occurrence of a storm S produced by atmospheric pressure A. Some basis must be found for excluding such non causal regularities. A favorite strategy of philosophers for accomplishing this appeals to the notion of "law." This is the strategy followed by Glennan (1996, 52), who characterizes mechanisms as follows: "A mechanism underlying a behavior is a complex system which produces that behavior by . . . the interaction of a number of parts according to direct causal laws." According to Glennan, the correlation between B and S is not a productive relationship because it is not a "direct causal law"; by contrast, relationships like (1) or, to use Glennan's example, the generalizations governing the interactions between the parts of a coke machine are "laws," albeit laws that are only "locally applicable" (1996, 15).

Like MDC, and in contrast to Glennan, I think that the notion of law is of very limited usefulness in characterizing the operation of most mechanisms. There is a major mismatch between the features that philosophers have thought laws must possess and the generalizations that characterize the operation of many mechanisms. Generalizations like Maxwell's equations and the field equations of General Relativity are paradigmatic laws; whatever else laws are, they are presumably generalizations that are relevantly similar to these paradigms. Laws are also commonly taken to be generalizations of wide scope that apply to many different kinds of systems and to have few or no (or at least a clearly delimitable set of) exceptions. In the above example, although the generalization $F = mg$ that describes the acceleration of a falling body is sometimes described as a law (the law of falling bodies), it lacks many of these features—it holds only approximately, even near the surface of the earth, and fails to hold even approximately at sufficiently large distances from the surface of the earth. It is obviously contingent on the earth having the particular mass and radius that it does. Generalization (1) is, if anything a less appealing candidate for a law. The textbook from which this example comes goes out of its way *not* to describe (1) as a law, but instead describes it as a non-fundamental and only approximate "empirical relationship" (Frautshi 1986, 179ff). The particular value of the coefficient of kinetic friction between a pair of surfaces is the net result of a large number of extremely complicated contact forces and depends, in ways that are still not well understood from the perspective of fundamental theory, on the detailed characteristics of the two surfaces and will change if the characteristics of those surfaces are altered. Thus, even if (1) correctly describes the frictional force on the block for some particular experimental set up, we could easily disrupt that generalization by, for example, greasing the surface of the plane or abrading it with sand paper. This is very different from the behavior of paradigmatic laws. A similar point holds for many other components of simple mechanical systems—springs, pulleys, gears and the like. The linear relationship between the extension and restoring force exerted by a spring depends on features that are highly specific to that sort of spring and will break down under sufficiently large extensions or other processes that deform the spring, the relationship between the movement of one gear and another will hold only so long as the gears remain relatively rigid bodies, and so on.

It is true enough that generalizations like (1) do possess at least one feature traditionally ascribed to laws: they support (some) counterfactuals—e. g., counterfactuals about how the frictional force will change if the normal force changes, as long as the experimental system is not disturbed by activities like greasing. However, without some further analysis of "what support for counterfactuals" means, this feature does not distinguish (1) from non-causal generalizations. The correlation between B and S also "supports counterfactuals," under one natural interpretation of that phrase: the counterfactuals "if the barometer reading is falling/ rising, the storm will /will not (or is more likely/less likely to occur)" are true, again so long as the system is undisturbed. What we would like is an analysis that gives us some insight into why one can appeal to generalization (1) but not the correlation between B and S in explaining the operation of a mechanism. Merely deciding to call the first but not the second a law does not help with this issue.

3. Interventions and Invariance. These remarks are intended to set the stage for the positive analysis that I favor. I will not repeat all of the details of that analysis here, since these are available elsewhere[1] but I will instead focus specifically on what my account implies for the understanding of mechanisms. First, what counts as regular "productive" behavior in a part or component of a mechanism? I understand this in terms of the notion

[1] See Woodward 2000.

of invariance under interventions. Suppose that X and Y are variables that can take at least two values. The notion of an intervention attempts to capture, in non-anthropomorphic language that makes no reference to notions like human agency, the conditions that would need to be met in an ideal experimental manipulation of X performed for the purpose of determining whether X causes Y. The intuitive idea is that an intervention on X with respect to Y is a change in the value of X that changes Y, if at all, only via a route that goes through X and not in some other way. This requires, among other things, that the intervention not be correlated with other causes of Y except for those causes of Y (if any) that are causally between X and Y and that the intervention not affect Y independently of X. Thus if A is a common cause of B and S as in the example above, manipulating B by manipulating A will not count as an intervention on B with respect to S since in this case the manipulation affects S via a route (the route that connects A to S) that does not go through B.

By contrast, if we were to employ a random number generator and, depending on its output, set the value of B at a high or low reading, in a way that is causally and probabilistically independent of the value of A, this would constitute an intervention on B with respect to S. What we expect of course is that the correlation between B and S would disappear if this intervention were to be performed repeatedly, which is to say that this correlation is not invariant under this intervention. Indeed, a moment's thought will show that the correlation between B and S is not invariant under *any* interventions on B. On the account that I favor, if a generalization G relating X to Y is to describe a causal relationship (or in MDC's language) a "productive" relationship, it must be invariant under at least some (but not necessarily all) interventions on X in the sense that G should continue to hold (or to hold approximately) under such interventions. The relationships described above meet this condition—for example, the relationship (1) will be invariant under at least a certain range of interventions that change the value of N—for some such changes, μ_k will be (at least approximately) a constant that is independent of N and hence (1) will correctly describe how F_k will change under this intervention. Similarly, the relationship $F = mg \sin ø$ will continue to correctly describe how the component of the weight of the block directed along the plane will change under interventions that change the value of ø and of m.

When a relationship is invariant under at least some interventions, it is potentially usable for purposes of manipulation and control—potentially usable in the sense that while it may not as a matter of fact be possible to carry out an intervention on X, it is nonetheless true that *if* an intervention on X were to occur, this would be a way of manipulating or controlling the value of Y. Thus, in the above examples, we may control the value of F_k by controlling the value of N, we may manipulate the restoring force exerted by a spring by manipulating its extension, and we may manipulate the component of the gravitational force on the block directed along the plane by manipulating m and ø. By contrast, one cannot manipulate whether a storm occurs by altering the position of a barometer dial.

This account seems to me to capture what is correct in the traditional idea that causal relationships, in contrast to non-causal relationships, support counterfactuals. What is distinctive about a generalization like (1) is that it supports counterfactuals of a special sort—counterfactuals that describe changes that will occur under interventions. The generalization describing the correlation between B and S may support other sorts of counterfactuals, but it does not support such interventionist counterfactuals. It is interventionist counterfactuals that tell us about the distinctive patterns of counterfactual dependence that are associated with manipulation and control.

The notion of invariance under interventions is intended to do the work (the work of distinguishing between causal and accidental generalizations) that is accomplished by the notion of law in accounts like Glennan's. As the above examples illustrate, whether or not a generalization is invariant is surprisingly independent of whether it satisfies many of the traditional criteria for lawfulness. For example, a generalization may be invariant (over some range of interventions and other sorts of changes) even though it holds only over a limited spatio-temporal interval or has exceptions or narrow scope or is not integrated into some larger body of theory. This is in large measure a consequence of the way in which invariance is defined: for a generalization to be invariant all that it is required is that it be stable under some changes and interventions. It is not required that it be invariant under all possible changes and interventions. Thus the generalization (1) describing the relationship between frictional and normal force is invariant as long as it would continue to hold under some interventions that change the value of N. Its invariance is not undermined by the fact that there are other interventions on N (for example, increasing N to a very large value) or other sorts of changes (greasing the contact surface) under which (1) would break down. Invariance thus has the virtue of capturing the idea that what really matters to whether a generalization describes a causal relationship is whether it describes a relationship that is potentially exploitable for purposes of manipulation and not whether it has the other features (wide scope etc.) traditionally assigned to laws. By contrast, someone who wishes to characterize (1) as a "law" is in the dialectically awkward position of needing to explain why that characterization is appropriate, even though (1) has exceptions, narrow scope, is not really well integrated with fundamental theory and so on.

4. Production and Counterfactual Dependence. My examples so far have involved simple physical systems. I turn now to a biological example which is designed to bring out the differences between the counterfactual account that I favor and a second approach that may seem attractive. This second approach attempts to look for some empirical feature or disjunction of features that underlies all cases of causal production. One candidate for this feature, emphasized in recent work by Wesley Salmon and by Phil Dowe is the transfer of some conserved quantity such as energy and momentum, perhaps in conformity with some continuity constraint. It may be that the list of "bottom out" activities in molecular biology proposed by MDC is advanced in a similar spirit—that their intent is to characterize empirically the notion of a "productive" interaction in molecular biology by providing a list of fundamental activities that count as productive.

Consider the lac operon model for E. coli due to Jacob and Monod. When lactose is present in its environment, E. coli produces enzymes that metabolize it. What is the mechanism that determines whether these enzymes are produced? According to the model proposed by Jacob and Monod, there are three structural genes that code for these enzymes as well as an operator region that controls the access of RNA polymerase to the structural genes. In the absence of lactose, a regulatory gene is active which produces a repressor protein which binds to the operator for the structural genes, thus preventing transcription. In the presence of lactose, allolactose, an isomer formed from lactose, binds to the repressor, inactivating it and thereby preventing it from repressing the operator, so that transcription proceeds. Biologists describe this as a case of "negative control." Unlike positive control in which, as the text I consulted puts it, "an inducer interacts directly with the genome to switch transcription on" (Griffiths 1996, 550) the inducer in this case, allolactose, initiates transcription by interfering with the operation of an agent that prevents transcription. The example is thus an instance of what has been called "double prevention" or "causation by disconnection" in the recent literature on causation (Hall, forthcoming).

What is interesting about this example, as well as other examples of double prevention, is that a causal relationship is present between the presence of allolactose and the production of the enzymes but there is no obvious respect in which there is transfer of energy from the former to the latter. And while we can perhaps use MDC's list of bottom out activities to describe the productive relationships between individual steps in the above process, it is far less obvious how to use this list to capture the idea that there is an overall productive relationship between allolactose and enzyme production without explicitly invoking the idea of counterfactual dependence. To begin with, the overall relationship between allolactose and enzyme production does not seem to fall into any of the categories on MDC's list. Nor is it plausible to claim that the overall relationship between X and Y is productive if X is connected to Y via a series of intermediate steps each of which correspond to an activity on MDC's list.[2] It also seems to me that the relationship between the presence of lactose and enzyme production as well as the relationship governing the intermediate steps of this process are too local and too susceptible to exceptions to count as plausible candidates for laws.

Nonetheless these relationships satisfy the condition described above for a relationship to be causal or productive: they are stable under some range of interventions. For example, the relationship between the presence of lactose and enzyme production is stable under interventions that change whether lactose is present, the relationship between the repressor protein and the operator is stable under interventions that change whether the repressor is present, and so on. These are also relationships that are potentially exploitable for purposes of manipulation and control—one can use them to manipulate whether enzymes are produced, whether the operator is repressed, and so on. In cases like this, a counterfactual theory of the sort outlined above seems to do a better job of capturing what production means than either a transfer of energy theory, a list of productive activities, or a law-based theory.

This idea that in molecular biology theories that describe causal or explanatory relationships or mechanisms provide information that is potentially relevant to manipulation or control is not just my conceit; it is explicitly endorsed by molecular biologists. A typical statement can be found in Robert Weinberg's (1985) discussion. He tells us that "[b]iology has traditionally been a descriptive science" but that because of recent advances, particularly in instrumentation and experimental technique, it is now appropriate to think of molecular biology as providing "explanations" and identifying "causal mechanisms." What does this contrast between description and the identification of causal mechanisms consist of? Weinberg explicitly links the ability of molecular biology to identify causal mechanisms with the fact that it provides information of a sort that could in principle be used for purposes of manipulation and control. According to Weinberg, biology is now an explanatory science because we have discovered theories and experimental techniques that provide information about how to intervene in and manipulate biological systems. Earlier biological theories—for example, traditional systems of classification of plants and animals—fail to provide such information and for this reason are merely descriptive. As Weinberg puts it, molecular biologist correctly think that "the invisible submicroscopic agents they study can *explain,* at

[2] This claim is not plausible because causation (and production) are not transitive. Standard counterexamples to the transitivity of causation, such as the dogbite example in McDermott 1995, are also counterexamples to the proposal in the text.

one essential level the complexity of life," because by manipulating those agents it is now "possible to *change* critical elements of the biological blueprint at will." (1985, 48, my emphasis). This account of the difference between descriptive and causal theories is essentially the view that I have defended above.

5. Modularity. So far I have been arguing that components of mechanisms should behave in accord with regularities that are invariant under interventions and support counterfactuals about what would happen in hypothetical experiments. However, I have said little about what a component is. The basic idea that I want to defend is that the components of a mechanism should be independent in the sense that it should be possible in principle to intervene to change or interfere with the behavior of one component without necessarily interfering with the behavior of others. As before, what is crucial is not whether such interventions can be carried out, as a practical matter, but rather whether this condition of independent changeability would be met *if* it were possible to carry out appropriate interventions on individual components. I will say that a system having this feature is *modular.*[3] Thus, if we imagine a machine consisting of components C_1.. C_n and some process that changes the behavior of C_1, then if the system is modular and we know what the new behavior of the changed component C_1 is, we can combine this with the generalizations governing the other components, which remain unchanged, to determine what the new overall behavior will be. This sort of independence of the various components allows one to trace out the consequences of possible changes in any of them for the overall behavior of the system. By contrast, if any change in (what seems to be) the generalization governing the behavior of one component automatically brings with it changes in the generalizations governing the behavior of (what we believe to be) other components, this is an indication that our proposed decomposition of the mechanism into parts or stages is incorrect.

As an illustration, return to (Ex1). As remarked above, it is natural to think of this machine as composed of two different components—a component force due to gravity directed down the incline of the plane and a force due to friction that resists the motion of the block. Applied to this example, modularity says that if these components are genuinely distinct, the relationships governing each should be independently changeable—that is, it should be possible to change the relationship or generalization governing the gravitational component without changing the relationship governing the frictional force and vice -versa. In fact, this is what we do

[3] For additional discussion of modularity in the context of systems of structural equations, see Woodward 1999.

find. We can change the relationship (1) by altering the characteristics of the contact surface between the block and the plane—e.g., by greasing it. This will change the coefficient of kinetic friction, but will not alter the generalization governing the gravitational force component at all. On the other hand, we could in principle alter the latter relationship by moving the inclined plane to a weaker gravitational field, which would result in a different value g' for the acceleration due to gravity. This would also change the value of N, the normal force exerted by the block, but it should not alter the relationship (1) between the frictional force and N. My suggestion is that at least part of what it means to say that we have identified the mechanism responsible for the overall motion of the block and that we have correctly segregated that mechanism into components is that we have exhibited that motion as the consequence of components that are independently changeable in the way just described. Obviously, once we have done this, we have the resources for answering a range of questions about how the motion of the block would have been different under changes in the inputs to these various components, as well as changes in the components themselves.

This idea yields the following proposal:

> **(MECH)** a necessary condition for a representation to be an acceptable model of a mechanism is that the representation (i) describe an organized or structured set of parts or components, where (ii) the behavior of each component is described by a generalization that is invariant under interventions, and where (iii) the generalizations governing each component are also independently changeable, and where (iv) the representation allows us to see how, in virtue of (i), (ii) and (iii), the overall output of the mechanism will vary under manipulation of the input to each component and changes in the components themselves.

My notion of modularity is closely connected to what Darden calls in her paper "modular subassembly" Recall that this is a strategy for mechanism discovery that proceeds by "hypothesiz[ing] that a mechanism consists of known modules or types of modules. One cobbles together different modules to construct a hypothesized mechanism" (Darden 2002). If this strategy is to work, it must be possible to add a new module or component to a structure consisting of other modules or to replace one module with another without disrupting or changing the other modules in the structure. This is just another way of expressing the idea that modules or components should be independently changeable. For example, in (Ex1) if we were to add an additional component consisting of a rope connecting a second weight to the sliding block via a pulley, this will result in an additional force on the block but it should not alter the previous two force components due to friction and the gravitational force on the block or the relationships (1) and (2) governing these. It is this that allows us to “cobble together” the original analysis of (Ex1) with the additional force component to produce an account of the new mechanism.

A similar modularity constraint holds for the biological example discussed above. In the case of the lac operon, the wild type gene I+ that produces the repressor protein can be replaced with a mutant form of the gene I− that does not produce the repressor. As expected, cells containing only I− but normal structural genes synthesize full levels of the enzymes in both the presence and the absence of an inducer. This is expected because, in accordance with modularity, it is assumed that this way of changing the repressor gene from I+ to I− does not affect the generalizations describing the operation of the structural genes, although it does of course cause the structural genes to be continually “on.” If cells contain both I− and a mutant form Z− of the structural gene that in its unmutated form (Z+) synthesizes B-galactosidase, they will not produce this enzyme. If an I+Z+ chromosome fragment is then introduced, recipient cells synthesize B-galactosidase for a period and then stop. The obvious interpretation is that this occurs because by the end of this period enough repressor product has been produced by the I+ genes to block further synthesis.

We thus see how the operon model in effect encodes a set of predictions about what will happen in various hypothetical experiments—experiments in which different components of the model: the repressor protein, the various structural genes, and so on—are manipulated independently of each other. If instead some other model of the mechanism was correct—if lactose induced enzyme synthesis by directly interacting with the structural genes (“positive control”)—then we would expect a different pattern of outcomes in these experiments. For example, we would not expect synthesis to occur in the presence of I− regardless of whether the inducer is present. An important part, then, of what it means to say that the operon model is a correct account of this mechanism and that it correctly describes the productive relationships at work in the mechanism is that it correctly predicts the results of these hypothetical experiments.

6. Conclusion. I believe that a notion of mechanism very similar to that characterized by **MECH** is employed in many other areas of science—for example, in psychology, where it is generally agreed that explaining the behavior of complex systems should proceed by decomposition into parts or modules each exhibiting a characteristic stable input/out relationship and where it is often explicitly assumed that the criterion for being a module or a part is that distinct parts must be “*separately modifiable*’” (Steinberg 1998, 706). However, while I will not try to argue for this thesis here, I believe that in contrast to the situation in mechanics and molecular biology, where purported mechanisms often do satisfy the conditions in **MECH,** there is frequently much less evidence that this is true in psychology. In psychology **MECH** has a real normative bite—in particular, the standard boxological diagrams allegedly describing the operation of psychological mechanisms drawn by psychologists are rarely accompanied by convincing evidence that the parts corresponding to the boxes satisfy the modularity condition described above. If the argument of this paper is correct, this is a reason for being skeptical that these diagrams describe genuine mechanisms.

REFERENCES

Darden, L. (2002), "Strategies for Discovering Mechanisms: Schema Instantiation, Modular Subassembly, Forward Chaining/Backtracking", *Philosophy of Science* 69 (suppl.):

Frautschi, S., R. Olenick, T. Apostol, and D. Goodstein (1986), *The Mechanical Universe: Mechanics and Heat.* Cambridge: Cambridge University Press.

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

Griffiths, A., J. Miller, D. Suzuki, R. Lewontin, andW. Gelbart (1996), *An Introduction to Genetic Analysis.* New York: W. H. Freeman.

Hall, N. (forthcoming), "Two Concepts of Causation."

Machamer, P., L. Darden, and C. Craver (2000), "Thinking about Mechanisms", *Philosophy of Science* 67: 1–25.

McDermott, M. (1995), "Redundant Causation", *The British Journal for the Philosophy of Science* 46: 523–544.

Steinberg, S. (1998), "Discovering Mental Processing Stages: The Method of Additive Factors", in D. Scarborough and S. Sternberg (eds.), *Methods, Models and Conceptual Issues: An Invitation to Cognitive Science.* Cambridge, Mass.: MIT Press, 703–863.

Weinberg, R. (1985), "The Molecules of Life", *Scientific American* 253 (4): 48–57.

Woodward, J (1999), "Causal Interpretation in Systems of Equations", *Synthese* 121: 199–257.

——— (2000), "Explanation and Invariance in the Special Sciences"*, The British Journal for the Philosophy of Science* 51: 197–254.