STUART S. GLENNAN

MECHANISMS AND THE NATURE OF CAUSATION

ABSTRACT. In this paper I offer an analysis of causation based upon a theory of mechanisms – complex systems whose "internal" parts interact to produce a system's "external" behavior. I argue that all but the fundamental laws of physics can be explained by reference to mechanisms. Mechanisms provide an epistemologically unproblematic way to explain the necessity which is often taken to distinguish laws from other generalizations. This account of necessity leads to a theory of causation according to which events are causally related when there is a mechanism that connects them. I present reasons why the lack of an account of fundamental physical causation does not undermine the mechanical account.

1. HUME'S PROBLEM

[E]xperience only teaches us, how one event constantly follows another, without instructing us in the secret connexion, which binds them together and renders them inseparable (Hume, 1777, p. 63).

Experience, according to Hume, cannot tell us about the "secret connexion" which binds together events. When we attend to a supposed causal interaction, for instance, a moving billiard ball colliding with a stationary billiard ball, we can observe the motion of the first ball and then the motion of the second, but we can not *observe* a connection between the two. Furthermore, no number of further observations would allow us to observe any connection.

Hume's problem is that, although we can observe regular conjunctions in nature, we can never see the "secret connexion" which binds them together. It would seem therefore that we can never have know what causes one event to follow another. His "skeptical solution" to this problem is to define the notion of cause in such a way that it makes no reference to a connection, but only to constant conjunction. He defines a cause to be "an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second" (Hume, 1777, p. 76). This definition is the germ of the regularity theory of causation which has dominated empiricist discussions of causation ever since.[1]

Although Hume's argument that no number of observations can yield an impression of a connection is, in my view at least, irrefutable, I do not

Erkenntnis **44:** 49–71, 1996.

think that that argument requires us to adopt a regularity view of causation. In this paper I will try to suggest an alternative that I call a mechanical theory of causation. The intuition behind the theory is straightforward. When I claim that some event causes another event, say that my turning the key causes my car to start, I do not believe this simply because I have routinely observed that turning the key is followed by the engine starting. I believe this because I believe that there is a mechanism that connects key-turning to engine-starting. I believe that the key closes a switch which causes the battery to turn the starter motor and so forth. Furthermore, this is not a "secret connexion". I can look under the hood and see how the mechanism works.[2]

There is an obvious objection to this sort of explanation. Although it works for cases like the key starting the car, there is a very important class of cases for which it does not. It is not possible, for instance, to look at the mechanism which causes two bodies to gravitationally attract each other. So far as we understand this interaction between bodies, there is no underlying mechanism which explains it. It is just a "brute fact" about the world in which we live. In this case we cannot explain the interaction by reference to any mechanism. Here some form of regularity theory seems plausible. The problem with the regularity theory is not that it is an incorrect analysis of such cases, but that it fails to distinguish these from cases where there is a discernible mechanism. The vast majority of cases are of the latter sort.

My aim in this paper is to offer an analysis of the concept of mechanism. At the conclusion of the paper I indicate how this analysis can be put to work in a theory of causation. I emphasize the fact that this theory cannot explain causation in fundamental physics. I suggest that there should be a dichotomy in our understanding of causation between the case of fundamental physics and that of other sciences (including much of physics itself).

Mechanical theories of causation are not new. The central tenet of the Mechanist movement in the seventeenth century was that all natural phenomena were explicable as the result of the action of mechanisms composed of corpuscles.[3] To say that some event caused another event was just to say that there is a mechanical linkage between the two, where this linkage was understood as a collection of corpuscles or larger rigid bodies pushing on each other. The failure of the Mechanist movement to provide an enduringly adequate account of causation stems, I think, from two problems – one scientific and one epistemological. The scientific problem has to do with the Mechanists' views about the microstructure of matter and the forces which govern interactions between matter. Mechanists advocat-

ed Democritean style theories in which the universe consists of a large (or infinite) collection of corpuscles which interact with each other only by collision (or perhaps more extended pushes). The admissible principles of interaction are what we would call "strictly mechanical principles", i.e., the sort of principles with which one would construct a mechanical (rather than say an electronic) device. It was believed that all apparently non-mechanical forces, such as gravitation and magnetism, could ultimately be explained as the actions of mechanisms consisting of corpuscles pushing on each other. This point of view has proved untenable in light of subsequent scientific developments. Physical forces such as gravitation and electromagnetism have resisted narrowly mechanical explanation.

The epistemological problem with seventeenth-century Mechanism has to do with the testability of mechanical theories. Mechanist explanations of natural phenomena are often dominated by baroque accounts of mechanisms that are not of discernible physical consequence (see, e.g., the discussion of gravity in Descartes, 1664). Although Descartes and other Mechanists claimed that all phenomena could be explained in terms of size, shape and motion of corpuscles, very little was said about how these corpuscles and their properties could be observed (or how we could make inferences about them on the basis of observables).

The account of mechanisms that I will develop is largely inspired by insights of the Mechanical philosophers. I hope, however, that my analysis will avoid these two pitfalls. In the first place, mechanisms must be conceived in such a way that there are not a priori restrictions on the sorts of allowable interactions which may take place between a mechanism's parts. Additionally, analysis of causal connections in terms of mechanisms is only meaningful when there are ways (even if indirect) of acquiring knowledge of their parts and the interactions between them.

In the remainder of the paper I will present an account of mechanisms and show how it provides the foundation for a theory of causation. Section 2 of this paper presents my analysis of mechanisms. Section 3 applies this analysis to two simple systems. Section 4 discusses some relationships between mechanisms and laws. Section 5 shows how the analysis of mechanisms can be used to formulate a theory of causation.

2. AN ANALYSIS OF MECHANISMS

There are two senses in which the term 'mechanism' is commonly used. The first sense refers narrowly to the internal works of machines, as when one speaks of a clock mechanism. The second refers more generally to complex systems analogous to machines, as when one speaks of a human

perceptual mechanism or a market mechanism. My analysis can be summarized by a defmition which is meant to capture this latter usage:

(M) A mechanism underlying a behavior is a complex system which produces that behavior by of the interaction of a number of parts according to direct causal laws.

Notice that (M) is a definition of a 'mechanism underlying a behavior' rather than a mechanism *simpliciter*. One cannot even identify a mechanism without saying what it is that the mechanism does. The boundaries of the system, the division of it into parts, and the relevant modes of interaction between these parts depend upon what the behavior we seek to explain. Furthermore, complex systems do many things at once. If one isolates a complex system by some kind of physical description, one can identify indefinitely many behaviors of that system (Kauffman, 1970). A complex system has many mechanisms underlying its different behaviors.

The polymorphous behavior of complex systems can be illustrated by considering the behaviors of the human body. Two of the many subsystems of the human body are the cardiovascular and respiratory systems. Each of these systems has mechanisms for doing certain things (pumping blood, inhaling oxygen and exhaling carbon-dioxide) and with regard to one of these (oxygenating blood) the two systems interact in such a way that they must be considered as a composite. These systems divide the body up in different ways. The cardiovascular system divides it into the heart, veins, arteries, capillaries, etc. The respiratory system divides it into lungs, diaphragm, windpipe, mouth, etc. The physical extensions of the systems and their parts overlap; there are, e.g., veins and arteries running through the various parts of the respiratory system. The choice of decomposition into parts depends upon the capacity or behavior to be explained (Wimsatt, 1972).

Although the choice of decomposition depends upon what is being explained, decompositions are not merely artifacts of the description. Veins and lungs are both really parts of human bodies, even though they overlap. Descriptions of mechanisms are good descriptions insofar as they describe what is "really" there. This point deserves emphasis because the context-dependence of systematic decomposition is often taken to imply anti-realism or relativism. Simple examples such as the one above show that there is no such implication.

The behavior in question may be something a mechanism was designed to do (or selected for), but it need not be. Consider two behaviors of a combustion engine: the motion of a drive shaft and the production of heat. Either of these behaviors may be legitimately mechanically explained.

However, the engine is designed to move the drive shaft, while the heat produced is merely a side-effect. When one considers designed artifacts or systems that have evolved under selection pressures, explanatory context often dictates analyzing the system in terms of its production of the designed or selected behavior, but one could choose any other behavior of these systems as well. In fact, one can investigate mechanisms which cannot in any interesting sense be said to have a purpose. One can consider, e.g., the solar system as a mechanism underlying the motions of the earth (or the planets) even though one believes that this motion is not purposive.[4]

In order for (M) to be sufficiently general it is important that a very wide variety of entities may be parts of mechanisms. Parts may be simple or complex in internal structure, they need not be spatially localizable, and they need not be describable in a purely physical vocabulary. In certain contexts, for instance, one might wish to consider genetic mechanisms whose parts are genes or information processing mechanisms whose parts are software modules or data structures. There are, however, certain kinds of entities which, to prevent (M) from being vacuous, should not be allowed to be parts of mechanisms. The parts of mechanisms must have a kind of robustness and reality apart from their place within that mechanism. It should in principle be possible to take the part out of the mechanism and consider its properties in another context. Care must be taken so that parts are neither merely properties of the system as a whole nor artifacts of the descriptional vocabulary. I shall summarize these restrictions by saying that parts must be objects.[5]

The significance of these restrictions can be illustrated by considering whether or not it is possible to give a mechanical explanation of the behavior of the electromagnetic field (as it is codified by Maxwell's equations). I believe it is not, because, in the only natural decomposition of the field into parts, the parts of the system are not objects in my sense. This case is significant because the electromagnetic field is an example of a law-governed entity whose behavior is not subject to mechanical explanation.

The electromagnetic field is an important part of many mechanisms, from particle accelerators to TVs to the mechanism which produces the Aurora Borealis. It is probably fair to say that electromagnetic fields play a role in producing the great majority of physical phenomena. Furthermore, electromagnetic fields are objects of a kind. They have a variety of properties – for instance energy and momentum. The properties of a field can be completely described by two vector fields, the electric field **E** and the magnetic field **B**.

Is there a mechanism that explains the properties of the electromagnetic field? If there is, then it should be possible to decompose the field itself into parts. There is at least one sense in which this can be done. It is quite appropriate to talk about the electric or magnetic field in a region – for instance, the electric field between two capacitor plates, or the magnetic field of the earth. These fields are parts of *the* electromagnetic field, if we conceive of this field as occupying the entirety of space. What justifies characterizing these parts as objects is their relatively high degree of separability from the field in surrounding space. However, this kind of articulation of parts would not be adequate to describe a mechanism that produces an electromagnetic field.

So far as I can see, the only possible articulation into parts consistent with the classical theory of the electromagnetic field involves thinking of the parts of the field as points in space. Each of these parts have the properties of electric and magnetic field strength. Such an analysis fails because points in space are not objects in my sense. Such points are not isolable or manipulable in themselves. It is not possible to differentiate experimentally between points in the electromagnetic field that are sufficiently close to one another. There are no boundaries between points in the way in which there are boundaries between, e.g., the field between the capacitor plates and the surrounding fields.[6] It is not possible to alter the electromagnetic field at a single point. The points, while part of the mathematical description of the field as a whole, have no physical significance apart from this description. It is crucial to a mechanical account that a system display some behavior which can be explained by reference to underlying properties of its constituents. But there are not, in this case, constituents with underlying properties. As Hertz remarked, "Maxwell's theory is Maxwell's equations".

The situation would be different if there were a detectable aether. If there were in any sense an underlying medium for the transmission of electrical and magnetic disturbances, then it might have been possible to investigate the properties of this medium. It might have been composed of particles and there might have been forces explaining how these particles interacted when disturbed by external sources (a light bulb or whatever). But according to our accepted physical theory there is no aether and there is no mechanical explanation of the electromagnetic field. One cannot then in any physically meaningful sense go deeper.[7]

The interactions between parts of mechanisms are, according to (M), governed by laws. I use the term 'law' here in essentially the same way as Goodman. Laws are generalizations (or universal propositions) which support counterfactuals. Lawlike or nomic generalizations are distinguished

from accidental generalizations because accidental generalizations offer no such support (Goodman, 1947; Nagel, 1961, ch. 4).

Some may argue that Goodman's account of laws is not adequate because it does not provide a criterion for demarcating "deep" laws (like Newton's law of universal gravitation) from less interesting counterfactual supporting generalizations (like whenever you leave bread on the counter for two weeks, it molds). Although it is true that Goodman's account does not provide such a criterion, for the purposes of my account of mechanisms, the lack of a clear distinction between these two sorts of counterfactual supporting generalizations is a virtue. The laws used to describe the interactions between parts of mechanisms and the laws which can be explained by mechanisms can be of the most profound or banal sort. In explaining a mechanism whose parts are interacting gravitationally we must invoke a rather deep law, while to explain mechanisms like lawn mowers we mostly invoke uninteresting counterfactual supporting generalizations, such as laws about the behavior of valves.

A more serious objection to my use of Goodman's analysis of laws is that it appeals to an unanalyzed notion of counterfactual support. Furthermore, it seems likely that any analysis of counterfactuals that we could give will appeal implicitly or explicitly to causal notions. Given that I intend to use my analysis of mechanisms to explain the nature of causation, this appeal threatens to make my analysis of causation circular. In Section 4 I will indicate how mechanisms can be appealed to in a non-circular manner to explain the notion of counterfactual support.

The final part of (M) that I wish to clarify is the stipulation that the laws governing interactions between parts of mechanisms be direct causal laws. The stipulation that the laws be causal is meant to exclude lawful generalizations which can be explained by common causes. For instance, it is a lawful generalization that night follows day, but certainly day does not cause night. Rather, the onset of day and of night are events which are both caused by the earth's rotation. Relations between parts must be governed by causal laws because otherwise the parts could not be said truly to interact.[8]

The further stipulation that causal laws be direct can be illustrated by the following example: Consider a system consisting of a series of three or more gears of various sizes. Given information about the number of teeth on the gears, one can state a law L_1 describing the rotation of the last gear as a function of the rotation of the first gear. L_1, however, does not describe an direct interaction between the first and last gear. The interaction is mediated by other parts which transmit the rotation from the first to the last gear. By contrast, a law L_2 describing the rotation of the second gear

as a function of the first is direct. What we are trying to capture by this stipulation is the sense that a mechanism is a collection of parts, in which the behavior of the aggregate stems from a series of local interactions between parts. In saying that an immediate interaction is local, I am not supposing that the interaction is spatially local, but only that there are no intervening parts. For instance, if gravitation were a genuine example of action at a distance, the gravitational interaction between the earth and the sun would be direct, because there would be no intervening parts. This has the consequence that the notion of directness is relativized to a particular decomposition (Glennan, 1992, ch. 3).

The analysis of mechanisms that I have given resembles a number of decompositional strategies for explanation, notably Kauffinan's articulation of parts explanation (Kauffman, 1970), Cummins' functional and morphological analysis (Cummins, 1975, 1980), and Haugeland's systematic analysis (Haugeland, 1978). There are, however, several important differences. First, each of these authors have emphasized the role of these explanatory strategies within a special science (psychology for Cummins and Haugeland, and biology for Kauffman), whereas my account is intended to apply to all sciences except fundamental physics. Second, I have tried to illustrate the connection between decompositional strategies and the explanatory role of mechanisms. Finally, and most importantly, I have (and will at the end of this paper) argued that these sorts of explanations are *causal* explanations, and more generally, that a relation between two events (other than fundamental physical events) is causal when and only when these events are connected in the appropriate way by a mechanism.

3. TWO SIMPLE MECHANISMS

In this section I will show how two simple mechanisms can be analyzed in the manner suggested by (M). The first example, a system to regulate the water level in a toilet tank, is clearly a mechanism by anyone's account. The principles (excepting gravity) according to which the parts interact are "strictly mechanical". The second example, a voltage switch, relies on different, "non mechanical" principles of operation. I will show that both of these systems may be analyzed along the lines of (M), and thus that they are mechanisms in my sense.

A. A Float Valve

Figure 1 pictures a simple mechanism to regulate the water level in a tank. It is called a float valve. It should be familiar to anyone who has ever opened the top of their toilet. Let us consider, in accordance with

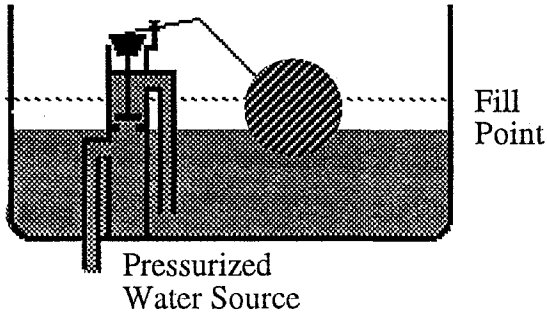

Fig. 1. A float valve.

(M), what the behavior of this mechanism is, what its parts are, and what causal interactions occur between these parts to produce the behavior of the mechanism. The purpose of a float valve is to regulate the water level of a tank, so it is natural (though not required) to single out the maintenance of a certain water level in the tank (the fill point) as the behavior of the mechanism. The operation of the mechanism is quite simple. A float is attached to a lever which opens and closes an intake valve. When the lever is down the intake valve is open, allowing pressurized water to fill the tank. When the lever is raised to a certain point, the intake valve closes, stopping the flow of water. The float is heavy enough that in the absence of water it will pull the lever down, opening the intake valve. On the other hand, it is sufficiently buoyant that the rising water level will force the float up, closing the intake valve when water reaches the fill point.

In describing the operation of this mechanism I have articulated a number of parts: the tank, the valve, the pressurized water source, the lever and the float. I have also specified the ways in which these parts are connected; that is to say, I have described the causal interactions between the parts. If one wished, one could formulate precise laws describing these interactions. One would do this by specifying properties of the parts, and using mathematical equations to represent these laws.

The description of the valve raises an important issue concerning the description of mechanisms generally. In the above description I have treated the valve as a kind of "black box" switch. It could be replaced by any device which allows water to flow into the tank only while the lever is in certain positions. I have not specified how the valve itself works, only how it contributes to the operation of the float valve mechanism as a whole. The accompanying diagram suggests some more detail. According to the diagram, the valve is a piston. As the lever is raised, the piston is lowered by a mechanical linkage. Since under normal operating conditions

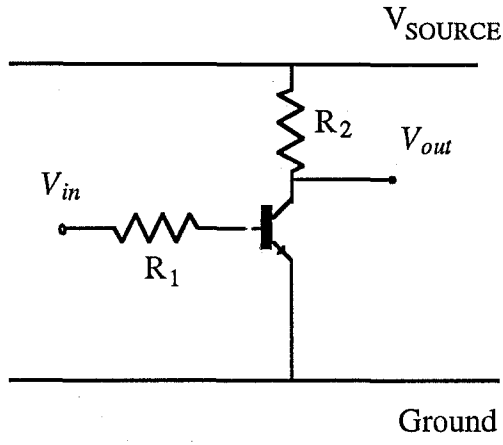

Fig. 2. A voltage switch.

the valve behaves according to simple laws, it is easy to treat it as a black box, abstracting away from the details of its operation. The valve is itself a mechanism within the larger water-level regulator mechanism. Though I have treated the valve as a simple part, one could equally well remove references to the valve, and replace it with references to the piston, the chamber, the lever-piston linkage, etc. This kind of reductive analysis is not limited to the valve. One could also, for instance, give an explanation of the details of the interactions between the water and the float; One could, so to speak, take the buoyancy mechanism out of its black box. In general, unless the laws of interaction between the parts of a mechanism are inexplicable in terms of any deeper physical mechanism, it will be possible to take apart the parts and look at how they themselves work.

B. A Voltage Switch

The float valve is a mechanism in the ordinary sense. The parts all interact by pushing each other. Water pushes float; float pushes lever; lever closes valve; water ceases to push. Though seventeenth century scientists would not understand the fluid mechanics as well as we do today, the qualitative principles were well known to them. The float valve would count as a mechanism by their lights. The voltage switch illustrated in Figure 2 operates according to quite different principles.[9] This is an electric rather than a mechanical switch. There are, at least at the macrophysical level, no moving parts. Nevertheless, it is easy to provide a mechanical analysis of this circuit analogous to the one provided for the float valve.

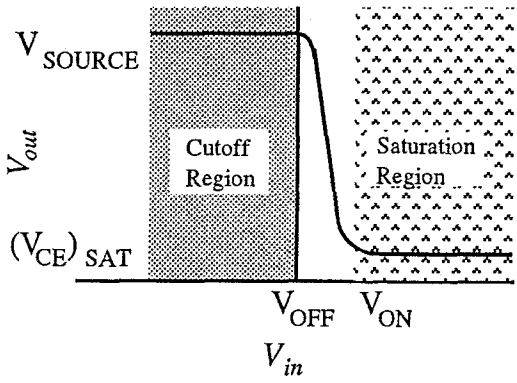

Fig. 3. Input/output behavior for a voltage switching circuit.

Already in describing the circuit in Figure 2 as a voltage switch, I have suggested a characterization of its behavior. The switch has two terminals, an input and an output. The behavior of importance in characterizing this circuit as a switch is the variation of input voltage V_{in} with respect to output voltage V_{out}. This behavior can be described quantitatively by giving a function $V_{out} = f(V_{in})$. For circuits of this type, the graph of this function has a distinctive shape illustrated in Figure 3. When V_{in} is below a certain voltage V_{OFF}, V_{out} is constant and equal to the voltage of the power source V_{SOURCE}. For values of V_{in} below V_{OFF} the transistor is said to be in the cutoff region. Then, for a certain interval beyond V_{OFF} (the transient region), V_{out} decreases approximately linearly as V_{in} increases. Finally, as V_{in} increases beyond the voltage V_{ON}, the output voltage levels out at a value near to 0, the collector-emitter saturation voltage $(V_{\text{CE}})_{\text{SAT}}$, and remains approximately constant for larger values of V_{in}. If $V_{in} \leq V_{\text{OFF}}$, the switch is off, indicated by $V_{out} = V_{\text{SOURCE}}$. If $V_{in} \geq V_{\text{ON}}$, the switch is on, indicated by $V_{out} = (V_{\text{CE}})_{\text{SAT}} \approx 0$.

To analyze the mechanism responsible for the switching behavior, we must now articulate the parts of the circuit. The central part is a junction transistor (hereafter, simply transistor). The transistor has three terminals: the base (left), the emitter (bottom) and the collector (top). The relevant properties of the transistor are given by the saturation voltage $(V_{\text{CE}})_{\text{SAT}}$ and a parameter β that determines the cutoff and saturation regions for the transistor. In addition there are two resistors, the bias resistor with resistance R_1 and the load resistor with resistance R_2. There is also a positive voltage source rail (top) with voltage V_a and a ground (bottom). These might be terminals of a battery. Finally there are input and output terminals. These parts are connected as indicated in the circuit diagram.

The circuit's behavior can be summarized by two simple equations. Define V_{OFF} to be 0 and V_{ON} to be $(R_1 V_{\text{SOURCE}}/\beta R_2)$. Then the equations are:

(1) $V_{in} \leq V_{\text{OFF}} \Rightarrow (V_{out} = V_{\text{SOURCE}})$ switch off

(2) $V_{in} \geq V_{\text{ON}} \Rightarrow (V_{out} \approx 0)$ switch on

The circuit is functionally a kind of current valve. When $V_{in} \leq V_{\text{OFF}}$ (i.e., voltage entering the base is negative), the valve is closed and no current passes from the emitter to the collector. When $V_{in} \geq V_{\text{ON}}$ (i.e., voltage entering the base is above a small positive value), the valve is open and current passes freely from the emitter to the collector. When $V_{\text{OFF}} < V_{in} < V_{\text{ON}}$, the valve is part way open, allowing restricted current flow between the emitter and the collector. Properties of the load resistor (R_2) and battery determine the voltage for the output terminal in the on and off states. When the valve is open, these properties also determine the output current $I_c (= V_a/R_2)$. Increasing the resistance of the bias resistor R_1 increases the voltage V_{ON} at which the valve opens.

This circuit illustrates a number of important features of mechanisms and how they can be analyzed in accordance with (M). Most importantly, this circuit is susceptible to mechanical analysis even though it is not, in the engineer's sense, a mechanical device. Additionally, as in the water regulator valve, the parts of the mechanism themselves are susceptible to mechanical analysis. It is possible to give mechanical explanations (in my sense) of the properties of the resistors, transistor, battery and conductors which indicate why they have the properties they do. Moreover, this circuit can itself be a part of larger mechanisms. Indeed, the chief interest of switching circuits is that they can be used as parts of larger electronic logic or control devices. Finally, this circuit indicates how what the mechanism is depends upon how you look at it. The description I have given of the circuit has been largely in the terms of elementary electronics. It would also be possible to give a description in terms of the microphysics of the system. This description would lead to a decomposition of the system in which the parts were electrons, molecular lattices, or other such entities. It would also be possible to consider this same circuit as a mechanism for other purposes. We could for instance consider the mechanism that produces heat in the bias resistor. Also, by arranging so that the input current always falls into the transient region, the same circuit can be used as a linear amplifier. The description given, though electronic, has emphasized the potential logical properties of the circuit. Other descriptions, appropriate to cases where input voltages typically lie within the transient region, will emphasize its linear behavior.[10]

The mechanisms that I have chosen to illustrate my analysis are both physical mechanisms. I have selected them to illustrate the point that (M) places no a priori restriction on the nature of physical interactions between parts. On the other hand, my analysis is in no way limited to mechanisms that are physical in nature. It is meant equally to apply to chemical, biological, psychological and other higher level mechanisms. I emphasize this point because the generality of the analysis is key if it is to provide a foundation for a theory of causation.[11]

4. FUNDAMENTAL AND MECHANICALLY EXPLICABLE LAWS

The behavior of a mechanism such as the voltage switch can be described in terms of one or more laws (i.e., counterfactual supporting generalizations) like those given in equations (1) and (2) above. A description of the internal structure of the mechanism explains this behavior. I call such laws mechanically explicable.

There is an important class of laws that are not mechanically explicable – fundamental laws. The essential feature of fundamental laws is that they are taken to represent facts about which no further explanation is possible.[12] While it is difficult to define the notion of fundamental law and arguably impossible to devise an adequate test to determine whether or not a law is fundamental, it is not hard to come up with a small body of laws which are nearly unanimously regarded as fundamental. Examples from classical physics include the law of universal gravitation and Maxwell's equations. In contemporary physics, this status is accorded to Einstein's equation relating mass distribution to space-time curvature or to Schrödinger equations for quantum mechanical systems. There are also many laws which, while scientifically quite significant, are not fundamental: for example, the ideal gas law, Hooke's law, and laws of classical genetics.

The claim for which I will argue is that all laws are either mechanically explicable or fundamental, *tertium non datur*. I will refer to this thesis as the thesis of the mechanical explicability of non-fundamental laws. In arguing for this thesis, we must be careful to construe it in such a way that it is neither obviously false nor trivial. A strong reading of it is that every instance of a particular lawlike regularity is explained by the operation of some particular type of mechanism. This reading would seem to entail that laws describing the behavior of higher level mechanisms would be type-reducible to laws of a lower level theory. But a number of widely accepted arguments (Davidson, 1970; Fodor, 1974; Kitcher, 1984;

Putnam, 1973; Wimsatt, 1976) suggest that type reductions of this sort are seldom possible.

A weaker reading of the thesis of the mechanical explicability of non-fundamental laws is that every instance of a non-fundamental law is explained by the behavior of *some* mechanism; but it need not be the case that the mechanisms which explain the various instances are all the same, or even of the same kind. For a higher level law to be mechanically explicable, it must be realized by some lower-level mechanism, but it may be multiply-realized. This reading would entail that laws describing the behavior of higher level mechanisms are token-reducible to lower-level laws, but not that they are type-reducible.[13]

The weak reading allows for the possibility that there are higher level laws, every instance of which must be explained by a different mechanism, perhaps even by a mechanism of a radically different kind. In such cases, the laws in question would not genuinely be *explained* by reference to these mechanisms, because nothing can be said about how the *type* of lawful behavior is produced by mechanisms. Such strongly irreducible laws would, like fundamental laws, resist mechanical explanation, but would, unlike fundamental laws, supervene on lower-level mechanisms.

On the weaker reading, the thesis of the mechanical explicability of non-fundamental laws is very plausible. The problem with the weaker reading is that it seems to entail little more than the claim that higher-level processes supervene on physical processes. However, insofar as I am committed to a particular analysis of mechanisms, the thesis says something about the way in which supervenience occurs, and in this regard it might be false. This point can be illustrated by the aether example discussed in Section 2 and footnote 7. If there were a genuinely continuous aether, we would say that electromagnetic properties supervene on properties of the aether, but we would not be able to say (without revising (M)) that the laws of electricity and magnetism were mechanically explicable. In addition, the thesis is not primarily a claim about the relation of higher-level laws to fundamental physical mechanisms, but rather a claim about the relation of higher-level laws to lower-level mechanisms generally. If one considers, for instance, the mechanism that explains Mendel's second law (the law of independent assortment of genes in gametes), the natural level of explanation is cytological. Mendel's second law holds (when it holds) because genes are often located on different chromosomes (or far away on the same chromosome), and given how meiosis works, a gamete is created by choosing "randomly" one chromosome from each chromosome pair. And while one could look at the mechanisms that explain meiosis,

the mechanisms that explain Mendel's second law are cytological, not physical.

A key feature of mechanically explicable laws is that there is an unproblematic way to understand the counterfactuals which they sustain. Consider a generalization about my car starting when I turn the key. I am justified in asserting "If I were to turn the key, the car would start" because I know that there is a mechanism which connects key-turning with car-starting. I also know the sorts of circumstances in which the counterfactual would turn out to be false, namely breakdown conditions for the mechanism which explains it. I know for instance that my key-turning would not lead to car-starting if the weather is too cold, or if there is no gas in the gas tank, because I understand the role of the battery and the gas in the ignition mechanism. Counterfactual generalizations can be understood in this way without appealing to unanalyzed notions of cause, propensity, possible world, or the like.

We are now in a position to partially address the worry raised in section II about whether the use of laws to describe the interactions of parts of mechanisms involves a circular appeal to causal notions. If the laws in question are mechanically explicable, then their lawlike (i.e., counterfactual-supporting) character can be explicated by reference to the further mechanisms which explain these laws. There remains the difficulty of understanding the source of lawlikeness for fundamental laws. I will not offer a solution to this last problem here, but I will argue in the final section that the absence of an account of the lawlikeness of fundamental laws does not undermine a mechanical analysis of higher-level laws and causal relations.

5. TOWARDS A MECHANICAL THEORY OF CAUSATION

I claimed at the outset of this paper that my theory of mechanisms could provide the foundation for a theory of causation. Although to spell out such a theory and defend it in any detail is beyond the scope of this paper, I can indicate briefly how such a theory would go, and suggest some consequences of the theory.

Before outlining such a theory, we should consider very generally the problems any theory of causation should solve. There are a number of standard ones: distinguishing real from spurious correlations; distinguishing lawlike from accidental generalizations; distinguishing real effects from artifacts. These are all instances of what we can, using Humean terminology, call the connection-conjunction problem. How does one distinguish connections from conjunctions? Humean approaches seek to solve the

problem by giving criteria that distinguish true from accidental regularities. Anti-Humean approaches typically appeal to some further notion of necessity (logical or natural) which distinguishes conjunctions from connections. I think that both of these approaches run into insurmountable difficulties, but I will not discuss these difficulties here.[14] Rather, I want to indicate how a mechanical theory tackles the connection-conjunction problem.

Roughly put, a mechanical theory of causation suggests that two events are causally connected when and only when there is a mechanism connecting them. How such a theory works is most clear in a case where the behavior can be described by a conditional. Take for instance the voltage switch discussed in Section 3. The behavior of the voltage switch can be summarized by three conditionals:

(1) $V_{in} \leq V_{\text{OFF}} \Rightarrow V_{out} = V_{\text{SOURCE}}$

(2) $V_{in} \geq V_{\text{ON}} \Rightarrow V_{out} = (V_{\text{CD}})_{\text{SAT}} \approx 0$

(4) $V_{\text{OFF}} < V_{in} < V_{\text{ON}} \Rightarrow (V_{\text{CE}})_{\text{SAT}} < V_{out} < V_{\text{SOURCE}}$

Because there is a mechanism which underlies this behavior, we can say, e.g., that increasing the input voltage from less than V_{OFF} to greater than V_{ON} *causes* a change in the output voltage from $(V_{\text{CE}})_{\text{SAT}}$ to V_{SOURCE}.

The chief virtue of the theory is that it makes the connection-conjunction problem a scientific one. If one can formulate and confirm a theory that postulates a mechanism connecting two events, then one has produced evidence that these events are causally connected. The necessity that distinguishes connections from accidental conjunctions is to be understood as deriving from a underlying mechanism, where the existence and nature of such a mechanism is open to empirical investigation. The mechanical account allows us to escape the regularity theory's difficulties with the connection-conjunction problem, while eschewing, as Hume did, reference to any metaphysical notion of necessity.

There is however an obvious limitation to the mechanical theory. Sooner or later the process of decomposition of a system into parts must come to an end. This is the level of fundamental laws. At this point we cannot point to any further or deeper mechanism. Since there is no mechanism, how do we explain the causal connection between events at the level of fundamental physics? The mechanical theory offers us no answers, and I will not try to add anything about the problem here.[15] For the moment, it is sufficient to recognize that whatever explains causal relations in fundamental physics is very different from that which explains causal relations at higher level.

There are two potentially serious objections to the theory that I have presented. The first of these alleges that any analysis of causation that

relies on mechanisms is circular, since any explication of the concept of mechanism requires the use of causal concepts. Of course one may identify a mechanism by articulating a system into parts and describing the behavior of the various parts; one may formulate statements describing the interactions between these parts and show how the behavior of the system as a whole (the effect) derives from the interactions between these parts. However, what does it mean to say that these parts interact? Is it not essential to the mechanical theory that changes in the properties of some parts *cause* changes in the properties of other parts?

This circularity is only apparent. In describing the mechanism that connects the two events I have explained how these events are causally connected. How the parts are connected is a different question. I can try to answer this second question by offering another account of the mechanisms which connect them, but I need not give such an account to explain the connection between the events. Indeed, such an account would only obscure the causally relevant features of the original explanation.[16] The supposed circularity is analogous to the apparent circularity involved in the recursive definitions of sentences of predicate logic. A typical clause of such a definition would be 'if p and q are sentences, then $(p \ \& \ q)$ is also a sentence'. Whether a string of symbols is a sentence depends upon whether certain other strings of symbols are sentences, but we are not offering a circular definition, because the sentences used in the definition can themselves be defined without reference to the sentence in whose definition they are being used. Similarly, in giving account of how two events are causally connected, I refer to a mechanism which in turn refers to causal relations, but these latter causal relations are different (and more basic) relations than the one which I am seeking to explain.

The second objection cannot so easily be defeated. Even granting that we can progressively explain interactions at one level in terms of mechanisms at the next, sooner or later, we are going to run out of levels and come to interactions governed by fundamental laws. I grant without argument that these fundamental interactions cannot be explained by the mechanical theory. But the fact that the mechanical theory gives us no account of such interactions combined with the fact that any mechanism depends ultimately on there being causal connections at the level of fundamental physics might lead us to believe that the mechanical theory has given us no account of causation at all. The objection is made eloquently by Hume:

> It is confessed, that the utmost effort of human reason is to reduce the principles, productive of natural phenomena, to a greater simplicity, and to resolve the many particular effects into a few general causes. ... But as to the causes of these causes, we should in vain attempt their discovery. ... The most perfect philosophy of a natural kind only staves off our ignorance a little longer. ... (Hume, 1777, pp. 30–31)

Certainly Hume is correct that there must be some facts which we cannot explain by reference to further more general principles (or mechanisms). Our explanations must stop somewhere. The question at issue is whether this ultimate dependence on unexplained regularities demands that we give up the mechanical theory and adopt a regularity theory of causation.

To understand why we are not forced to adopt a regularity theory, we must look at how it is that we in fact evaluate the truth of causal claims. Although there are many ways to do this, I submit that the best way to evaluate such claims is to find the mechanism responsible for the supposed causal connection. If for instance, we want to show that smoking causes cancer, the best way to do so would be to discover the mechanism by which tar, nicotine, etc., interact with the body to produce cancerous cells. We might provide overwhelming statistical evidence to show the correlation between smoking and cancer, but so long as we do not understand the mechanism in question, we can still wonder whether or not the correlation indicates that smoking *causes* cancer.[17] Not only are regularities insufficient to establish causal connections, they are unnecessary as well. Once we have identified the mechanism, we need not acquire additional evidence for the regularities it produces. Also, further detail about the nature or operations of the parts of the mechanism are not relevant. The best way to find out if it is a dead battery that is preventing my car from starting is to use a voltmeter to test the battery's charge. Once I have established that the battery has no charge, I have sufficiently confirmed my hypothesis, and taking apart the battery will provide no additional evidence. Although the mechanism responsible for connecting two events may supervene upon other lower-level mechanisms, and ultimately on mechanically inexplicable laws of fundamental physics, it is not these laws which make the causal claim true; rather, it is the structure of the higher level mechanism and the properties of its parts.

To illustrate some of the benefits of a point of view which treats causation in fundamental physics differently from higher level causation, I would like briefly to sketch how the point of view of the mechanical theory can shed light on an interpretive problem in the quantum theory. It is often said that the quantum theory, while extraordinarily successful as a predictive instrument, cannot be said to explain the phenomena that it predicts. This predicament comes out most clearly in the case of the "unexplainable" correlations produced by EPR type experiments.[18] The problem may be illustrated as follows. It is possible to construct a device which shoots a pair of particles in opposite directions to distant targets. These particles can be prepared in such a way that, upon hitting these targets they will deflect in one of two directions, up or down. It is not possible using this

preparation technique to determine in advance which direction they will go. However, quantum mechanics predicts and experiment confirms that if one particle deflects up then the other particle will deflect down and vice versa.

There seem to be two explanations which might account for this correlation. First, there might be some signal sent from one target to the other. This possibility is ruled out (if relativity theory is right) by placing the targets so far apart that any interactions would require signals to travel faster than the speed of light. Alternatively, one might think that the preparation of the particle pair puts the particles into a certain state which causes them to go in one direction or the other when they reach the target. Surprisingly, however, mathematical results (so called no-hidden-variable-theory results) indicate that there can be no such state.

This result is generally considered to be very strange and hard to understand. It indicates (consistently with experimental evidence) that there are correlations between events where neither a direct causal connection nor an indirect connection via a common-cause can possibly account for that correlation. I think that the analysis of causation that I have offered shows why we should not be so surprised. What is so puzzling about the correlation in question is that there is no mechanism which could possibly connect the events occurring at the two targets and that there is no mechanism which could possibly connect them each to some third common cause. However, if one believes that quantum mechanical laws describe the most fundamental physics, then one does not believe that there is a deeper mechanism anyway. And if there is no such mechanism, what reason is there to believe that distal events should or should not be correlated? Our uneasiness derives in part because we expect that the laws describing this quantum mechanical system should have properties similar to those of mechanically explicable laws; but there is no reason to have such an expectation. It is an artifact of our belief that there is something behind the regularities.

The mechanical theory of causation rejects a wide-spread assumption about the nature of causation. I think that it is generally assumed that whatever causal connections are, they ultimately have something to do with the most fundamental physical processes. The closer we are to fundamental physics, the more our statements are about the true causes of things; the further we stray into the higher-level sciences, the more we move away from causal statements and toward mere empirical generalizations. This assumption, however, is what makes Hume's skepticism so devastating. On this assumption causes are the ultimate metaphysical glue which holds fundamental physical events together. Hume provides a convincing argument

that we can have no knowledge of this glue, and that talk of such glue may even be unintelligible. The solution to these difficulties is to reverse the initial assumption. Causal statements are typically statements about events regulated by mechanisms, and mechanisms are complex, higher level entities. Only when we talk about interactions governed by fundamental laws does causal talk become problematic.

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

NOTES

* I would like to thank Erich Reck, Mike Price, Bill Wimsatt, Ron McClamrock, Dan Garber, Howard Stein and Ken Waters for fruitful discussions and for comments on earlier drafts of this paper.

[1] I do not wish in this introduction to make claims about the proper interpretation of Hume. In particular, I am ignoring the following question: Does Hume's skepticism lead him to believe that there are no powers behind regularities, or only that these powers are unknowable? Although twentieth century empiricists tend to argue for the former interpretation, the textual evidence is not decisive. For a defense of the latter view, see Strawson (1989). Whatever the correct interpretation of Hume, it is sufficient for my purposes that Hume's argument is often taken as leading to the regularity theory of causation.

[2] I should emphasize that I do not personally need to have any knowledge of what is under the hood in order to make a causal claim. It is sufficient that I believe that there is something under the hood which experts understand or which is open to empirical investigation.

[3] For a discussion of Mechanism and Corpuscularism in the seventeenth century see Glennan (1992, ch. 2).

[4] 4 Since (M) ignores etiological or teleological constraints, it makes very few restrictions on what could count as a mechanism. For instance, it is quite possible to describe my belt as a mechanism for stopping bullets. Generally a mechanism such as this is not worthy of investigation, but we can imagine contexts where it would be. If I was saved from an untimely death because a bullet ricocheted off of my buckle then this mechanism would not seem so silly.

[5] I do not mean here to give a definitive analysis of the notion of object. I do, however, think that my use of the term is plausible, because whatever objects are, it seems to be important that they can exist in a variety of contexts.

One problem with this stipulation is that there are certain entities which I would like to consider as objects which are in a natural sense properties. For instance, beliefs and desires (or mental states generally) are often described as properties, but I want to allow for mechanisms in which such entities are parts. There is nothing, however, that prevents us from viewing such entities as objects under some descriptions and as properties under others. Under one description beliefs and desires are properties of our brains, while under another description (if some brand of cognitivism is right), beliefs and desires are themselves objects which have various computational properties.

[6] In the case of a physical "system" like the electromagnetic field, the requirement of empirical isolability is roughly the requirement that the parts be discrete. I have used the more general term because in cases where parts are not spatially localizable it would not be clear to what a requirement of discreteness amounted.

[7] The possibility that the electromagnetic field could be explained in terms of properties of aether raises a further question about (M). A detectable aether would provide an explanation of properties of the electromagnetic field, but would it then be correct to say that aether is the mechanism that transmits electromagnetic waves? If (M) is correct, then it would only be appropriate to call the aether a mechanism if it were possible to decompose it into discrete parts. But it is possible that the aether would turn out to be a genuinely continuous medium, and if this were so no such decomposition would be available. Supposing that there were a detectable aether, there would be a further empirical question whether or not the aether was composed of discrete particles.

In short, the possibility of a genuinely continuous aether raises a question about the generality of (M). (M) defines mechanisms in such a way that all mechanisms are collections of discrete (in the sense of decomposible) parts, but a continuous aether would be a sort of "continuous mechanism". If we wished to count the aether as a mechanism, we would then have to amend the definition (M) to allow for "continuous" mechanisms having a single continuous part. Currently, our best physical theories do not demand such emendation.

[8] If this account is ultimately to work, one must have an adequate way to distinguish between these two kinds of laws. One of the most promising approaches to solving this problem is by use of the statistical relationship known as "screening off" (see e.g., Salmon, 1984). For a discussion of the limitations of this approach together with alternatives, see Glennan (1992, ch. 5).

[9] This example is taken directly from an undergraduate electronics text (Calvert and McCausland, 1978). See §9.4 and §14.2.

[10] It is the higher level description (e.g., whether the circuit is an amplifier or a switch), chosen by considering the context in which the circuit occurs, that allows us to determine what aspects of the lower-level behavior are significant as opposed to noise. For a similar point see McClamrock (1995) Chapter 1.

[11] Examples from sciences other than physics are discussed in Glennan (1992). Chapter 6 contains a case study of two models of vowel normalization mechanisms that have been developed by cognitive psychologists.

[12] I am presuming that these laws are laws of physics.

[13] It would take a fuller exposition to spell out the relationship between different interpretations of mechanical explicability and various types of reducibility. As framed by Fodor (1974), type and token reducibility are theses about the reducibility of theoretical terms and laws of one theory to theoretical terms and laws of another. While there is an analogous reduction relations between laws and the mechanisms which realize them, some work must be done to show how type-reducibility of laws and theoretical terms is related to type-reducibility of mechanisms.

[14] I suggest some reasons in Chapter 8 of Glennan (1992). See also Salmon (1984) for an excellent discussion of why regularity theories cannot solve the connection-conjunction problem. In that book (Chapters 5, 6 and 9), Salmon proposes an account of causation that is meant to address this problem and that in certain ways parallels my own account. Space does not permit me to discuss Salmon's theory in detail, so I can only mention one problem which suggests that his theory is incomplete. Salmon's theory is concerned with distinguishing causal processes from what he calls pseudo-processes". Pseudo-processes are the sorts of things which produce non-causal but lawlike regularities. Salmon's criterion for distinguishing between these processes is that causal processes can transmit "marks". Salmon has unfortunately not explained why causal processes transmit marks where pseudo-processes do not. The true difference between causal processes and pseudoprocesses can only be explained, in my view, by considering the differences in the mechanisms underlying them.
[15] I discuss how to explicate fundamental causal relations in a way that dovetails with the mechanical account in Glennan (1995).
[16] Kitcher (1984) makes a similar observation in his discussion of his thesis R_3.
[17] A similar point has been made in the case of evolutionary biology by Sober and Lewontin (1982). They argue that the existence of actual or dispositional regularities between events (or properties) is not sufficient for attributing a causal relationship between those events (or properties). They conclude from this that regularity accounts of causation are inadequate. The mechanical theory of causation I have presented allows one to distinguish between causal and artifactual correlations of this sort.
[18] For a discussion of EPR correlations see, e.g., Shimony (1989).

REFERENCES

Calvert, J. M. and M. A. H. McCausland: 1978, *Electronics* Vol. X of *The Manchester Physics Series*, F. Mandl, R. J. Ellison and D. J. Sandiford (eds.), John Wiley & Sons, Chichester.

Cummins, Robert: 1975, 'Functional Analysis', *Journal of Philosophy* **72**, 741–765.

Cummins, Robert: 1980, *The Nature of Psychological Explanation*, MIT Press, Cambridge.

Davidson, Donald: 1970, 'Mental Events', in L. Foster and J. W. Swanson (eds.), *Experience and Theory*, University of Massachusetts Press, Amherst.

Descartes, René: 1664, *Le Monde, ou Traité de la Lumière*, edited and translated by Michael S. Mahoney, Abaris Books, New York, 1979.

Fodor, Jerry A: 1974, 'Special Sciences, or The Disunity of Sciences as a Working Hypothesis', *Synthese* **28**, 97–115.

Glennan, Stuart S.: 1992, *Mechanisms, Models and Causation*, Ph.D. Dissertation, the University of Chicago.

Glennan, Stuart S.: 1995, 'A Two-Tiered Theory of Causation', unpublished manuscript, Butler University

Haugeland, John: 1978, 'The Nature and Plausibility of Cognitivism', *Behavioral and Brain Sciences* **1**, 215–226.

Hume, David: 1777, *Enquiries concerning Human Understanding and concerning the Principles of Morals*, reprinted from the 1777 edition, 3rd ed., P. H. Nidditch, Oxford University Press, Oxford, 1975.

Kauffman, Stuart: 1970, 'Articulation of Parts Explanation in Biology and the Rational Search for Them', in R. C. Buck and R. S. Cohen (eds.), *Boston Studies in the Philosophy of Science*, PSA 1970, Vol. VIII, D. Reidel, Dordrecht.

Kitcher, Philip: 1984, '1953 and All That. A Tale of Two Sciences', *Philosophical Review* **93**, 335–74.

McClamrock, R.: 1995, *Existential Cognition: Computational Minds in the World*, University of Chicago Press, Chicago.

Nagel, Ernest: 1961, *The Structure of Science*, Harcourt, Brace Jovanovich, Inc., New York.

Putnam, Hilary: 1973, 'Reductionism and the Nature of Psychology', *Cognition* **2**, 131–146.

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

Shimony, Abner: 1989, 'Conceptual Foundations of Quantum Mechanics', in Paul Davies (ed.), *The New Physics*, Cambridge University Press, Cambridge.

Sober, Elliott, and Richard C. Lewontin: 1982, 'Artifact, Cause and Genic Selection', *Philosophy of Science* **49**, 157–80.

Strawson, Galen: 1989, *The Secret Connexion*, Clarendon Press, Oxford.

Wimsatt, William C: 1972, 'Teleology and the Logical Structure of Function Statements', *Studies in History and Philosophy of Science* **3**, 1–80.

Wimsatt, William C: 1976, 'Reductionism, Levels of Organization, and the Mind-Body Problem', in G. G. Globus, G. Maxwell and I. Savodnik (eds.), *Consciousness and the Brain: A Scientific and Philosophical Inquiry*, Plenum, New York.

Manuscript submitted October 10, 1994
Final version received July 13, 1995

Dept. of Philosophy and Religious Studies
Butler University
4600 Sunset Ave.
Indianapolis, IN 46220
U.S.A.