Behavior and Philosophy, *32*, 149-166 (2004). © 2004 Cambridge Center for Behavioral Studies

CAUSAL THEORIES OF INTENTIONAL BEHAVIOR AND WAYWARD CAUSAL CHAINS

Berent Enç
University of Wisconsin—Madison

ABSTRACT: On a causal theory of rational behavior, behavior is just a causal consequence of the reasons an actor has. One of the difficulties with this theory has been the possibility of the "wayward causal chains," according to which reasons can cause the expected output, but in such an unusual way that the output is clearly not intentional. The inability to find a general way of excluding these wayward chains without implicitly appealing to elements incompatible with a pure causal account (like brute acts of will) has been a problem for the causal theory. This essay attempts to find a general solution to the problem. The solution rests on the premise that behavior-producing systems are goal-directed, and that on a purely causal analysis of goal-directedness it can be shown that the wayward chains' resulting in the goal is purely fortuitous because these chains do not subserve the function of the system.

Key words: causal theory of behavior, wayward casual chains, naturalized teleology

Philosophical literature concerning human behavior can be separated into two camps. The one camp, which we may call, for lack of a better label, the Cartesian Camp, views human action as sui generis, an irreducible capacity of human agency. The second camp starts with the presumption that human action is part and parcel of the natural order of things and can be fully understood in terms of the causal connections among events in nature. Thus what separate these camps from each other are irreconcilable doctrinal differences.

According to the first camp, humans are *essentially* capable of *initiating* actions, either as bodily movements or as mental acts of willing or intending. In this approach the notions of initiating action, voluntary control, and the power to determine one's action are taken to be part of a family of notions that includes agency, and it is taken to be a conceptual truth that members of this family cannot be analyzed in terms of notions outside the family.

On the view held by the second camp, which we may label the naturalist view of action, human action is simply the output of a rather complex organism in response to the inputs and the current states of the system. Here the force of the word "response" is to be explicated exhaustively in causal terms.

To many who see science as ultimately illuminating all aspects of the universe, the naturalist approach seems the only sensible one—but this essay is not

AUTHOR'S NOTE: This essay is a shortened and revised version of a chapter of my forthcoming book *How We Act: Causes, Reasons, and Intentions* (Oxford University Press). I am indebted to remarks on an earlier draft by Ellery Eells, Larry Shapiro, and Elliott Sober. I am also grateful to José Burgos and François Tonneau for their corrections and helpful suggestions.

an examination of the pros and cons of these two approaches. I find myself firmly placed in the naturalist camp, and I will not try to argue against the Cartesians here. The purpose of this essay is to take one criticism that is typically directed at a naturalist view of human action and attempt to show that the criticism fails.

Let me first do some stage-setting.

Ordinary ways of talking about human behavior allow us to distinguish between our deliberate intentional *actions* and things that we do instinctively in response to things that happen to us. This distinction is sometimes marked by separating *action* from *mere behavior*. Confronting a dangerous situation, we sometimes quickly deliberate the pros and cons of several courses of action, choose one, and act on the choice. We then act intentionally—this is full-fledged action—but sometimes fear takes hold of us and "makes us" run by instinct. The commonsensical view is that under these circumstances we are not acting as rational agents; our behavior cannot be explained in terms of any *reasons* that we may be said to have for running.

When a naturalist, one thinks of human action as the output of a system that is the causal consequence of certain parameters, and one finds oneself confronted with a choice.

The first alternative is to dismiss the commonsensical distinction and refuse to treat behavior due to the reasons the agent has as a separate category from other kinds of behavior. The second alternative is to try to accommodate the distinction in the differences between two types of causal antecedents. In this way one would then confront the task of naturalizing the notions of agency, of initiating action, and of voluntary control. Philosophical tradition has opted for this second alternative. Myles Brand (1984) has characterized the attempt to find a mark of difference between rational action and mere arational behavior as the Fundamental Problem of Action Theory. Since Brand many naturalizers of action have taken this problem seriously, and I tend to agree with the tradition.

On a superficial glance, the problem might seem rather easily solved. After all, in describing the difference we were able to speculate, quite plausibly, that in the case of action the output is the causal consequence of a state reached by a deliberative process, which may be thought of as a computation of the anticipated costs and benefits of different alternative outputs, whereas in the case of the so-called mere behavior we can suggest that some input gets processed though pathways that shunt out the cognitive and other intentional states of the system, and this process results in the output. The computation that precedes action clearly involves the beliefs and desires or preferences of the agent. There is the general presumption among naturalizers, however, that these intentional states can be shown to be amenable to naturalistic explication. If this presumption is justified, the schema outlined here seems to solve the Fundamental Problem of Action theory simply and neatly.

To put this thought in more schematic terms, we may offer the following definition of action—a definition that is designed to mark action off from "mere behavior."

Schema A: S’s bringing about an event E is an action if and only if the occurrence of some mental state (M) in S causes S to bring about E.

The choice of M is clearly a key move in fleshing out this definition. Because the focus of this essay is elsewhere, I will simply stipulate that the occurrence of M is the formation of the intention to act in a certain way, and the intended act is to be described as the bringing about of the event E. A necessary condition for a state to count as an intention to do something will be that the state was reached as a result of some computational process that weighed the consequences of the alternatives.

We can illustrate the way Schema A works by a pair of examples:

(E1) John is a rock climber leading one other climber. At one point John realizes that if he holds onto the rope that supports the second climber they will both fall and die. He decides to let the rope go to save his own life, and as a result of this decision he releases his hold on the rope. The second climber falls and dies, and John is saved.

(E2) Joe finds himself in exactly the same situation as John, but before Joe can think through his predicament panic overtakes him, causing his grip on the rope to slacken with the same outcome.

Clearly, John acted intentionally and Joe did not. Schema A gives the right answer and explains why. John’s act of letting the rope go was a direct consequence of the formation of his intention to let the rope go, whereas Joe never formed such an intention.

So far so good. But we can see how a problem can be raised for Schema A. Suppose a third scenario, as described by Donald Davidson (1980), in which Sam, who is in the same situation as John and Joe, starts deliberating along the same lines as John. Just like John, Sam arrives at the intention to let the rope go, but the intention fills Sam with such horror that it makes his hands spasm, and consequently the rope is released with the familiar result. Now, despite the fact that the intention to let the rope go caused the release of the rope, it is intuitively clear that Sam’s behavior was not an intentional action, so Schema A fails to provide a sufficient condition for full-fledged (intentional) action.

This inadequacy in Schema A is generally recognized, and it is revised to read as:

Schema B: S’s bringing about an event E is an action if and only if the occurrence of some mental state M in S causes S to bring about E *in the right way*.

The task then reduces to that of specifying what the “right way” is. Appearances notwithstanding, this task has proved to be a difficult one, and the literature is full of attempts to give a general characterization of this right way. The attempts take the form of giving a set of necessary and sufficient conditions that define this right way, but for each attempt counterexamples are concocted that involve causal pathways that satisfy the proposed conditions yet clearly yield nonintentional behavior. Such causal pathways have been called wayward (or deviant) causal chains. The ease with which these counterexamples are found and the pretheoretical intuitions that enable us to separate out actions from mere behavior without any reference to event causation have encouraged Cartesians to maintain that no general naturalist account can be given for this right way. They have argued that this fact by itself is enough to undermine naturalism in action theory—naturalism understood as the thesis that necessary and sufficient conditions for what counts as an intentional act can be formulated in terms of the causal connections among mere events.[1]

The problem here is a general one. In defining action, Schema B proposes to separate actions from mere behavior not in terms of the intrinsic properties of the bodily outputs but in terms of their etiology. In that way Schema B falls in the general category of the so-called philosophical causal theories. Causal theories of knowledge, perception, and representation are some examples. Because each theory has the same structure as that of Schema B, it has to confront the task of saying what the right way is, finding some criterion that will distinguish between the normal causal chains and the so-called wayward chains. Some theorists bracket the task and others confront it squarely. I tend to think that a general account of what constitutes the right way in *all* causal theories is possible, but here I shall confine my remarks to the causal theory of action. I propose to proceed as follows.

I shall first look at a type of attempted solutions to the problem of wayward chains. I will offer a general argument as to why that type is vulnerable to counterexamples, then I will use the argument as a stepping stone to formulating a different type of solution. I will defend this solution by showing how it gives the right answer for different kinds of examples of wayward chains discussed in the literature. If the solution is successful, one major obstacle to naturalizing action will have been overcome.

Before I begin I should introduce one final preliminary notion because some later distinctions I want to draw will make better sense if I can avail myself of this notion.

In action theory it is sometimes assumed that there is a beginning point to action—that there is something we can identify as the *first thing* that we do. I turn on the light by flipping the switch. Flipping the switch is an intentional act of mine *by means of which* I turn on the light. It might also be true that I flip the switch *by* doing something else intentionally, namely by moving my finger. But I do not move my finger by doing any intentional act. I certainly move my finger by sending nerve impulses to my muscles, but sending nerve impulses to my muscles arguably is *not* an intentional act of mine. This pattern of thinking allows us to identify moving my finger as a *basic act*. Defining basic acts is notoriously difficult, and I will not attempt it here. A vague notion of something we can do "readily," without having to figure out how to do it, is all I need for my purposes. I will assume that we all have a repertoire of basic behaviors, some of which, like walking or playing a tune on the piano, are complex units that in some contexts are

[1] One of the earliest sources of this argument is found in Taylor (1966, p. 249). More recently, Moya (1990) used such an argument to dismiss causal theories of action. Davidson, too, who, in my mind has given the best reasons for subscribing to a causal theory of action, confesses to being defeated by this problem. I tend to think that Davidson's arguments for the thesis of the Anomaly of the Mental rely in an essential way on his view that if the problem of wayward causal chains cannot be solved, then all attempts to have a genuine science of psychology are bound to fail.

produced as basic intentional acts and in other contexts (e.g., teaching someone to play the tune) are made up of smaller units of basic intentional acts (like pressing the individual keys).

I can now begin.

Perhaps one of the earliest and the most careful attempts comes from Peacocke's (1979) general treatment of causal waywardness. He imposes three conditions that must be satisfied by the cause–effect relations if these relations are to be nonwayward. He takes these conditions to spell out the requirement that there be some *sensitivity* relation between the cause and the effect. (1) Differential explanation: when the effect's having some property is explained by the cause's having some property, the explanation should be supported by a law that includes a mathematical functional relation between the degree to which the cause has that property and the degree to which the effect has that property (p. 66). (2) Recoverability: the cause should be necessary for the effect.[2] (3) Stepwise recoverability: the requirement of (2) should apply to each pair of consecutive links of the causal chain (p. 80). These requirements amount to demanding that if the chain is to be nonwayward, then the following counterfactual conditional must be true: if the cause had been different, then each of the subsequent links in the chain would show a corresponding difference.

One feature of Peacocke's way of capturing the sensitivity relation is that the three conditions offered in the definition of this relation impose constraints on the structure of the causal path that connects the cause with the effect. They suggest that in the causal theory of action, for the chain to be normal, the behavioral output caused by the intention must be *sensitive* to the content of the intention in such a way that incremental changes in the content of the intention are accompanied by corresponding changes in each of the links ending with the behavior. This requirement is, or course, violated in the example of the third rock climber, Sam, whose intention to let go of the rope so unnerves him that it causes him to loosen his hold. Here we cannot find a smooth mathematical function that relates some feature of the content of the intention to the corresponding features of any of the intermediate links of the relevant causal chain.

It can be shown, however, that this specific version of the structural requirement is too strong. An example of Frankfurt's (1969) that has come to be known as the case of the counterfactual contravener can be put to good use here. In this example we are to imagine that there is a neuroscientist who would interfere and make me do A if I were to fail to intend to do A. Despite this, when I intend to do A and I do A as a causal consequence of my intention, my doing A is clearly intentional. However, the counterfactual reading of the requirement that if I had intended to do something other than A, then I would have done that other thing is violated—whatever the intention, the output will always be the same behavior. When the contravener is inactive, however, the way the intention to do A causes

[2] *Op. cit.*, pp. 79-80. As Peacocke puts it, if *p* differentially explains *q*, then given just the initial conditions and *q* one shall be able to recover *p*.

the behavior A is perfectly normal, hence the behavior is intentional. Thus, the example shows that Peacocke's conditions are too strong.

This is perhaps not too serious a problem. These conditions can be revised or weakened to accommodate this type of counterexample, and this is exactly what has happened in the literature.[3] I shall not pursue this line here, for I take a second, more general problem to be more telling. I think it is possible to scaffold a general argument to show that *any* requirement formulated in terms of the properties of the structure of the causal chain from the intention to the behavior has to fall short of diagnosing the real source of waywardness.

Toward this end, let me take an example of a wayward causal chain. Suppose that an actor has been criticized in the past for not doing nervous scenes well. On the opening night of a new play, because his part requires it, he intends to appear nervous, yet his intention causes him to *be* nervous, and as a result of his nervousness he ends up appearing nervous. I think it is clear that, at least on the opening night, his appearing nervous was not intentional. The scenario, in its relevant features, is not significantly different from that of the third rock climber, Sam.

To start the argument, let us first suppose that there is some structural feature of the causal chains that normally lead from the intention to the behavior. The feature is such that when it is absent the resulting behavior is not intentional, even when it is caused by the right intention. Let us assume that the requirement for nonwaywardness is formulated in terms of this structural feature. I will call such a requirement R. In the nervous actor example the behavior's being nonintentional was caused by the fact that the causal chain leading from the intention to the behavior violated R.

Let us now entertain a second version of the nervous actor scenario. In this second scenario the actor can exploit the way in which R was violated in the original scenario and incorporate it into his action plan. He will know that just forming the intention to appear nervous will result in his thinking of delivering his lines as if he were nervous, whereupon his personality will make him nervous, and this will yield the intended result.[4] In this case the actor's appearing nervous will be caused by his intention to be nervous, and the causal path will be *identical* to the causal path of the first scenario, yet in this second version he will be appearing nervous intentionally. One and the same path is wayward in the first scenario and normal in the second.

This pair of examples suggests a general argument. Whenever we are given a theory of nonwaywardness that imposes the requirement R to be satisfied by the

[3] A very comprehensive treatment of waywardness and an elaborate set of conditions designed to capture what counts as an intentional action can be found in Mele & Moser (1994). For other examples see Audi (1973), Bach (1978), Brand (1984), Davies (1963), Gibbons (2001), Mele (1992), Searle (1983), and Tuomela (1977).

[4] The basic act in the act plan that supports the second scenario is *thinking of* delivering the lines as if he were nervous, but this difference between the scenarios is not reflected in the causal path, The same events are linked in the causal chain in both scenarios; it is just that in scenario 1 his appearing nervous is a basic act type, which gets tokened by a behavior that is not a basic *act* token, whereas in scenario 2 appearing nervous is an *intentional* nonbasic act.

structure of the causal pathway, we can find two systems that have identical causal pathways that violate R, one of which is wayward and the other of which is normal. This general argument exploits the intuition that when a system contains a wayward causal path it achieves what it is *supposed to do*, but not *in the way* it is supposed to do it. The argument proceeds by showing that for some system doing what it is supposed to do by violating R is not that system's doing it in the way it is supposed to do it, whereas for another system doing that same thing by again violating R *is* that second system's doing it in the way it is supposed to do it. Conversely, a system's operating in the way it is supposed to by conforming to R does not guarantee that a different system that also conforms to R will operate in the way it is supposed to operate. In summary, *the way* a system is supposed to do something is not capturable by stipulating certain structural requirements on the causal path.

The proper response to the general argument is to turn to the *system* in which the causal chain is being evaluated and develop a theory of what should count as a normal chain *relative* to the well functioning of the system.[5]

Typically, the causal theories one encounters in the literature have a common feature—they all involve a system that has the function of bringing about a result or a final product. The normal path that connects the input to the final product is one that subserves the function of the system. More specifically, when causal theories offer a causal analysis of a type of entity, the type of entity is the output produced by the system *when and only when* the system's producing that output is part of what counts as the system's executing its function. In other words, the "right," "ordinary," or "characteristic" ways in which the required type of event causes the output are nothing other than the ways in which the system executes its function. On the other hand, if a token causal chain assumes a path that is not "normal," the system might end up doing what its function is to do without actually *functioning* at all, that is, it does "what it is supposed to do," but not "*in the way* it is supposed to do it."

Perhaps the easiest way to illustrate this is to take artifacts that are designed to produce outputs that represent certain features of their environment. For example, the thermometer is a device that has the function of representing the room temperature, and it succeeds in executing its function *only when* the causal path through which the temperature changes in the room bring about changes in the pointer position is the same as the causal path that was anticipated in the design of the device. This representational relation *is severed* if a change in the room temperature causes a corresponding change in the pointer position through a *different* causal path. As an example, we may imagine a case in which the room temperature takes a rapid fall from 20° Celsius to 0° Celsius, which causes a crack in the wall. The thermometer falls off the wall, its pointer jumping to the 0° mark, and gets stuck there. Such a different path interferes with the functioning of the device: it causes it to malfunction or it causes it not to execute its function. That is perhaps the clearest example of a case in which a system does what it is supposed

[5] The importance of teleological considerations for waywardness was anticipated by Davies (1963).

to do (have its pointer reading correspond to the room temperature) but not *in the way* it is supposed to do it. What a system is supposed to do is determined by its function, and the way it is supposed to do it is determined by its design. It is important to keep the class of cases in which the thermometer fails to (correctly) indicate the room temperature (i.e., fails to do what it is supposed to do) separate from the class of cases under discussion. In the former class of cases the pointer reading fails to indicate what the room temperature is just because the room temperature is *not* what the reading says it is. This can happen in one of two ways: (1) because the thermometer is malfunctioning (e.g., the pointer is stuck) or (2) because although the thermometer is operating exactly as it was designed to operate, the environment does not cooperate (e.g., the thermometer is placed, by mistake, on top of a radiator). The issue of distinguishing wayward causal chains from normal ones does not arise in this class of cases simply because the distinction between wayward and normal is made against the background supposition that the system in question *is* doing what it is supposed to do. However, just as a system can fail to do what its function is to do in one of two ways (by malfunctioning or by the environment not cooperating), a causal chain that subserves a function can, in general, end up being wayward in one of two ways: by the system's malfunctioning or by the environment's providing *haphazard* cooperation. In the first way, despite its malfunctioning, the system fortuitously brings about the expected result. In the second way, the system might not be malfunctioning, yet outside the system a sequence of events might occur, and these events might, by luck, "cancel each other" in such a way as to render the chain wayward. The account I propose to develop starts with the basic intuition that a key to understanding what counts as a "normal" causal chain is the concept of executing a function—the concept of a system's doing something that it is supposed to do *in the way* it is supposed to do it. When we move away from designed artifacts to natural systems these concepts become harder to apply.

Here one can avail oneself of a common move of understanding the "design" of natural systems in terms of adaptation by natural selection, and this is what I propose to do. Flowers achieve cross-pollination by emitting scents that attract bees to their stamen. If the scent offends me and I toss the flower out in the wind, cross-pollination is affected, but not *in the way it is supposed to be*. The flower was not "designed" to have the scent to cause the pollen to be carried to a pistil in this way. Here, the notion of how things are *supposed to be* and *design* are metaphors that can be cashed out in terms of the evolutionary history of the flowers and bees. A similar story can be told for the way the mammal eye functions and science fiction scenarios can be concocted to introduce wayward chains into the perceptual process.

I offered the conjecture above that the key to understanding waywardness in causal chains is to look at the well functioning of the systems in which the chains subserve some function of the system. If this conjecture is correct, the logic of function assignments will help us to develop an adequate requirement for what counts as normal chains in causal theories.

On one account of functions, function assignments to a system can be analyzed by the following clauses:[6]

The function of a system (S) is F if and only if

(i) In response to the onset of a certain set of specifiable conditions (C), S produces output (O),

(ii) O results in F, and

(iii) The causal chain from C to O to F is such that for any intermediate link (X) in the chain, the fact that C causes X is explainable by the fact that X causes F.

In the flower briefly mentioned above, the fact that the internal state of the plant causes the fragrance is explainable by the fact that the fragrance attracts the bees. This explanation is provided by the theory of Natural Selection. Natural Selection has brought it about that there is a nomic dependence of the fact that the plant produces this fragrance on the fact that the fragrance attracts the bees.[7]

On the other hand, if on one occasion I get offended by the fragrance and throw the flower away, thereby causing cross-pollination, the fact that the plant produces the fragrance cannot be explained by the fact that the fragrance offends me—until the time when my behavior begins to contribute causally to the prevalence of the fragrance in the population of these flowers—that is, until the fragrance becomes an adaptation for its capacity to elicit this type of behavior from me. This is just to reiterate the obvious fact that selection produces not just traits but also the processes that lead to the expression of these traits.

Needless to say, Natural Selection is just one of the ways in which the explanatory condition (iii) above can be satisfied. Another way in which the dependence between the causal connection from C to X and that from X to F can be established is through Skinnerian operant conditioning. A third way is through deliberative reasoning.[8]

Returning to action theory, we can reword the causal theory of action more explicitly as follows:

(CTA) The behavioral output of an organism is an intentional action (A) if it is caused in the way it is supposed to be caused by an intention to do A.

(W_o) An intention to do A causes a behavioral output in the way it is supposed to if and only if for any intermediate link (X) from the intention to the behavior, the fact that the intention causes X is explained by the fact that X results in that behavior.[9]

[6] See Enç (1979). For the purposes of the problem of deviance, many other naturalistic analyses of function (e.g., Millikan, 1989, just to mention one out of many) will be compatible with these claims.

[7] There are close enough possible worlds in which the bees fail to be attracted by the fragrance. Most of these are worlds in which the plant has gone extinct or the fragrance has atrophied. It is true that among them there might be possible worlds in which the bees are not attracted to the fragrance, yet the fragrance is retained because it is genetically linked to some other feature that is functional or because, in terms of ecological economy, it is too "expensive" for the plant to eliminate the mechanism responsible for the production of the fragrance. The fact that these latter worlds are possible does not affect the truth of the claim that the required explanatory relation holds.

[8] See Enç & Adams (1992) for a discussion of the affinity between functions and purposive behavior.

[9] To be more general, instead of saying ". . .is explained by the fact that X *results* in the behavior," I should say ". . .is explained by the fact that X *makes it the case that* the behavior emerges" because,

In the formula (W_o), all the links in the chain, including the intention to do A, are assumed to be events. The fact that the forming of an intention is a *mental* rather than a *physical* occurrence need not give us pause here because we are operating with the naturalistic presumption that all mental events are at bottom certain neurophysiological changes in the brain.

To capture the force of (W_o), the formula can be more perspicuously stated as follows. In any causal chain, starting with the intention and terminating in a behavior, we can speak of the cause–effect pairs in one of two ways. From one perspective these pairs are event *types* that can be subsumed under causal laws, either deterministic or probabilistic. From a second perspective the same pairs are temporally fixed specific event *tokens*. When the agent forms the intention at t_0 to release his hold on the rope, this event causes some specific event X at t_1—for example, a neural signal being sent to the muscles, and X, in turn, causes the token relaxation of the hand muscles at t_2—the behavior in question. In addition, there is a regularity that governs these causal relations. Intention to release one's hold does, other things being equal, cause events of type X, and events of type X do (again other things being equal) cause the relaxing of the hand muscles. I intend the formula to be understood as requiring that the regularity between the intermediate event *type* X and the *type* of behavior in question is what does the explaining. To capture this lawful regularity I will word the requirement by saying that the fact that X *would* generate the behavior must explain why the intention caused X. Also, although it does not make much difference in the examples discussed so far, the requirement will stipulate that what is *explained* be the causal relation between the *token* intention and the *token* X.

(W) An intention to do A causes a behavioral output in the way it is supposed to if and only if for any intermediate link (X) from the intention to the behavior, the fact that a tokening of that intention causes a token X is explained by the fact that under the circumstances, that type of X would generate that kind of behavior.

Armed with this final version of the formula for waywardness, we can now look at some typical examples of wayward chains from the literature and see how well the formula fares. These examples are traditionally divided into two subcategories, one in which waywardness infects the causal chain from the intention to the basic act (sometimes called "antecedential waywardness"), and the other in which the causal chain from the basic act to the distal event brought about by some nonbasic act is wayward (sometimes called "consequential waywardness").[10] This difference between the two subcategories is reflected in the view I defend here. When we examine the causal pathway from the intention to the action in the framework of explanations derived from the function of the system we can distinguish cases in which the system malfunctions (the first subcategory of antecedentially deviant cases) from those in which the system functions well but the cooperation of the environment is totally fortuitous (the second subcategory of consequentially deviant cases). I shall argue that both types of cases are covered by

as we will see below, sometimes the relation between X and the behavior will be constitutive (i.e., they will stand to each other in what has been called "a generative relation").

[10] These terms, "antecedential" and "consequential," were introduced by Myles Brand (1984).

the same requirement, W, of the explanatory relation between the two segments of the causal chain.

The example of the nervous actor has the same pattern as that of the third mountain climber, Sam. The intention to appear nervous causes (via the thought of appearing nervous) a state of nervousness, and the state of nervousness makes the actor appear nervous. We may say that the function of the intention-generating system is to produce actions that match what is intended, so we expect from nonwayward operations of this system that for any intermediate stage in the chain from the intention to the basic act, the fact that the intention produces this intermediate stage is explainable by the fact that this intermediate stage causes the basic act. (The explanation is to be provided by looking to the "design" of the system and seeing that each intermediate stage has been "chosen" by natural selection for its capacity to contribute to the production of that output which the system has the function of producing.) In the scenario in question, however, it is clear that the fact that the intention causes nervousness has nothing to do with the fact that the nervousness results in appearing nervous. The two facts are independent of each other. Given the details of the scenario, the intention to appear nervous would have caused (via the same path) the nervousness *even if* (due to possible self-control on the part of the actor) the nervousness did not have the propensity to result in appearing nervous.

When we move to the second, nonwayward version of the scenario, in which the actor exploits the wayward causal path of the first scenario and in doing so transforms his appearing nervous into an intentional act, we can see that the explanatory requirement is satisfied.

In this version the revision we made in the earlier version of the formula W_o helps us see the importance of the type-token distinction. In applying W we ask "Is the fact that the actor's intention to appear nervous (according to the plan described) causes him to think of delivering his lines as if he were nervous explained by the fact that thinking this way would cause him to appear nervous?"

The answer is clearly "yes." The whole "point" of the actor's intentionally generating this thought in his mind is that he knows that the thought will make him nervous, thus securing the event of his appearing nervous.

Several well-known examples have been offered as cases of consequential waywardness. Perhaps the classical one is Chisholm's (1966) example of the rich uncle. Carl intends to kill his uncle to inherit his fortune. While he is driving to his uncle's house, where he intends to commit the act, his thoughts agitate him and make him drive carelessly. As a result he hits and kills a pedestrian, who, as luck will have it, happens to be his uncle.

Other examples involve a different kind of luck. In one, an inept sheriff wants to shoot a robber. He takes aim, but he is a terrible shot. The bullet goes in the wrong direction, hits a spittoon, ricochets, and kills the robber.[11]

Philosophers who have looked at wayward causal chains in action have given one kind of treatment for antecedential wayward chains and a different kind for

[11] See Brand (1984, p. 18).

consequential chains. One point in favor of formula W is that it uniformly gives the right answer, not only for those two kinds of waywardness but also for two additional kinds I shall discuss later.

When we examine the case of Carl and the rich uncle, we can clearly see that Carl's driving carelessly at the time of the accident is not explainable by the fact that his driving carelessly was to cause his uncle's death. The two facts are independent of each other. The fact that Carl's driving carelessly would cause the death made no contribution to the fact that his intention made him drive carelessly.

Again, when we look at the case of the inept sheriff, we can conclude that the intention's causing the gun to be pointing in the direction of the spittoon when the trigger was pulled is not explainable by the fact that the gun's being pointed in that direction would cause the robber to be shot. The sheriff's ending up with the gun pointing in that direction when he pulled the trigger had nothing to do with the fact that pointing the gun in that direction would be causally sufficient for the death of the robber.

It might be instructive here to contrast the case of the inept sheriff with a case in which a man intends to kill his victim by shooting him in the heart. He takes aim at the heart and kills his victim by shooting him in the head. The intuition is that whereas the sheriff's killing the robber was not intentional, the man in this example killed his victim intentionally. I think that the judgment depends on how the intention is fleshed out. In the normal case in which one aims at the chest to kill someone, one realizes that aiming at the chest might result in hitting him in the head, thus killing him. Therefore, a case can be made that part of the reason why the man's intention resulted in his aiming at the chest is that aiming at the chest would result (given certain other factors, which might obtain) in hitting him in the head, thereby resulting in his death. For instance, it might be true that he aimed at the chest *rather than lower* because aiming at the chest gives him a better chance at hitting him in the head, which is very likely lethal.

However, changing these details might make the example wayward, and the explanatory requirement W would explain why. Suppose that the man believes the victim is coated in armor save one hole over his heart. Unbeknownst to him, there is also a hole in the armor covering his head. If he aims at the heart, thinking this is the only way to kill him, but the wind blows the bullet with the result that it hits him in the head, then it seems that the fact that (given the characteristics of the wind) aiming at the heart would result in the victim's death (by hitting him in the head) does *not* explain why his intention resulted in his aiming at the heart. What explains why his intention resulted in aiming at the heart is the fact that in normal conditions (when there is no such wind) aiming at the heart would result in killing him by hitting him in the heart. In this example this latter law is not the operative law that explains why in this token case aiming at the heart killed the victim, so we do not have the right explanatory relationship between the two segments of the causal chain. This supports the intuition that the man killed his victim nonintentionally.[12]

A third kind of luck is exploited in concocting cases of waywardness that have been labeled "tertiary waywardness."

Here are two examples:

(i) Catherine intends to enter the room after the queen enters. The queen has already entered when Catherine sees Anne enter the room; she mistakes Anne for the queen and enters after Anne. It was just luck that the queen had entered before Anne, so Catherine does not intentionally enter the room after the queen.

(ii) Al is taking a multiple-choice test. The test has a question sheet and an answer sheet. In the question sheet four answers, (a) through (d), are offered, and Al is supposed to pick the correct one and fill out the letter that corresponds to the correct answer on the answer sheet. On one question the possible answers are (a) bee, (b) ant, (c) spider, (d) scorpion. The correct answer is "ant." Al, however, thinks the correct answer is (a). He circles (a) in the question sheet, and when he starts to transcribe his answer on the answer sheet, he fills out the letter (b). (Here the coincidental match between "bee" and "b" is what makes the chain go wayward.) He intended to provide the right answer and his intention caused him to circle the right answer, but he did not provide the right answer intentionally.[13]

In the case of Catherine, it was just fortuitous that Anne had entered the room after the queen. So it is *false* that the fact that Catherine's intention caused her to enter after Anne is explainable by the fact that her entering after Anne would constitute her entering after the queen (i.e., would make it the case that she entered after the queen). The connection between Catherine's entering after Anne and her entering after the queen is due to chance, so the fact that her entering after Anne *was* an act of entering after the queen cannot explain why Catherine's intention caused her to enter after Anne. The explanatory requirement is violated, and we do not have an intentional act.

It is important to be sensitive to the causally relevant properties of the events being considered to see if the explanatory relation holds or not. The property of Catherine's basic act that was relevant to that act's being an act of entering after the queen was its being an act of entering after Anne. That property of her basic act was not caused by Catherine's intention to enter after the queen. The explanatory relation is severed when the intention's causing some event is independent of that event's bringing about the satisfaction of the intention.

The case of Al is not significantly different. His intention to provide the right answer caused his circling (b) in the answer sheet, but this is not explainable by the fact that his circling (b) would constitute his circling the right answer "ant." What explains this is his belief that the right answer was "bee" plus a simple parapraxis.

[12] A similar pair of examples were used by Mele & Moser (1994, pp. 50-51). They also concluded that although the first killing was intentional, the second was nonintentional because, according to them, it diverged from the plan the agent had formed in executing the intention.

[13] The term *tertiary* was coined by Mele (1987) as applying to this example.

Finally, there is a fourth family of cases that are not typically raised in discussions of wayward causal chains in the theory of action.[14] These are cases in which the causal pathway leads from the intention of an agent via the action of a mediating second agent to a behavior of the first. Some of the causal pathways result in an intentional action; others render the act nonintentional.

I think the best way of approaching this family of cases is first to look at prosthetic devices.

A(1) It is, I think, absolutely clear that whether one has a prosthetic arm or a real arm makes no difference to our judgments about the intentionality of an act that involves the basic act of an arm movement, and even when the prosthetic device has a loose wire that makes contact only some of the time, when it does make contact and the arm moves, the movement would be intentional.

A(2) Suppose the prosthesis is further upstream in the causal pathway. Suppose, for example, that when the intention to raise the arm is formed, a brain state, B, results, and electrodes in the brain transmit by radio signals the information about B to a satellite computer, which in turn sends signals to the electric motor of an artificial arm that moves the arm in the way intended. Again, the intuition is that the agent raised the arm intentionally. This is borne out by the fact that, given the design, the system that includes the agent and the elaborate prosthesis was functioning in the way it was supposed to, and the explanatory requirement that tests for this gives the right answer: the fact that the radio signal would cause the arm to rise explains why the intention causes the radio signal. The engineers who designed the prosthesis would not have incorporated a circuit in which the intention causes the signal if the signal could not get to produce the intended arm movement.

A(3) If the prosthetic setup in A(2) is functioning only intermittently, then when it does work the arm raising should be judged as intentional. This judgment parallels the case in which a person is struck by paralysis 90% of the time she intends to move a limb, yet when she does move the limb as intended the movement is judged intentional. When the prosthesis works, the system is functioning as it is supposed to and there is no room for waywardness. As I argued above, the lack of reliability in the workings of the prosthetic device is not enough to produce waywardness.

In cases A(1) through A(3) we have been assuming that the agent is not aware of the existence of the prosthetic device. This is realistic enough because most persons are not aware of the neurophysiological causal pathways that lead from their intention to their basic acts, so I have not been implicitly relying on a belief condition involving how the arm moves. If we give the agent the knowledge of the working of the device, the only thing we change *might* be that the arm movement is no longer a basic act.

A(4) Suppose, however, that there is no design, no "device," no system to function or malfunction. Bishop and Peacocke mention an example of David Pears.' In the example, a gunman's nerve to his finger is severed, but his brain

[14] Bishop (1989) and Ginet (1990) are notable exceptions.

event caused by the intention to pull the trigger attracts a lightning bolt, which provides the required impulse to the finger.[15] The consensus is that the gunman pulls the trigger intentionally.[16] I have to disagree. My appeal to a system that at the time of the action was performing its function in the way it was supposed to is inapplicable here, hence the explanatory requirement is violated. I admit that some readers might regard this case as a refutation of the solution I have been defending here, but the following three-step reasoning might challenge the intuitions, whatever they may be worth, in such fantastic science-fiction cases:

(A) Sally buys a lottery ticket with the intention to win. When the improbable happens and she does win, the intuition seems to be clear that it is *false* to say that she won the lottery intentionally.

(B) In a game of darts I aim at the bull's eye and throw the dart. When the improbable happens (let us say that due to my ineptness at the game the probabilities are comparable to those of Sally's ticket being the winner) and I do hit the bull's eye, it seems equally clear that I hit it intentionally.

What is the difference between (A) and (B)? One conjecture I can offer is that in the dart example the result depended *on the way I threw the dart*. Although it was a "one in a million" affair, at that moment my system functioned the way it was supposed to. But Sally's method of picking the ticket made no contribution to the fact that the ticket was a winner. When I made the successful throw, I, by luck, tapped into a system that had specific well-functioning conditions. But in Sally's case there was no such system, hence the fact that the ticket she picked would be the winning ticket cannot explain why she picked that ticket.

(C) Fred shoots an electron gun. There are 100 locations the electron can randomly land in. One of the locations is rigged to a device that will execute a cat. In the remaining 99 positions the cat will survive. Fred wants the cat to die, and he is lucky; the electron lands in the one slot that results in the cat's death. It is true that he killed the cat.[17] But did he kill it intentionally? The intuitions might urge us to say "yes," but if my conjecture about the difference between (A) and (B) is right, then (C) is clearly of the same kind as winning the lottery because there is no system of which it can be correctly said that it was functioning the way it was supposed to, so I submit that in (C) our intuitions mislead us. Because it involves a gun we mistakenly subsume it under acts of "aiming" and ignore the fact that the way Fred pulled the gun made no contribution to the result. Admittedly, all of these examples involve nonbasic acts, but they illustrate how intuitions can be shaped by mistaken assumptions. My defense in the lightning bolt example is that we arrive at the wrong intuition, that the act is intentional because we subsume the random atmospheric occurrence under designed prosthetic aids.

Using these judgments we can now move to examples involving two agents.

[15] See Pears (1975).

[16] Bishop (1989, p. 174, footnote 14) and Peacocke (1979, pp. 89-93) share this intuition. Pears originally disagreed on the grounds that the process is not reliable, but he is reported by Peacocke to have changed his mind later. I do not hold reliability of the process to be a necessary condition for nondeviance; nonetheless, I am committed to judging this to be a deviant path.

[17] The example is borrowed from Dretske & Snyder (1972).

B(1) Suppose that some connection in a prosthetic device is severed, and a second agent holds the wires together. Clearly, this addition to the scenario does not affect our original judgments.

B(2) If the second agent of B(1) had a shaky hand and the connection is secured only some of the time, still the agent equipped with the prosthesis acts intentionally whenever the connection happens to be made. This case is no different than that of the intermittent prosthetic device in A(3).

B(3) Suppose, again, that an agent is equipped with a prosthetic device and that a second agent has some way of telling (e.g., by means of some advanced use of f-MRI) what the first agent intends to do. The prosthetic device is disconnected from the first agent's brain, but the second agent sends a signal through his computer to the prosthetic device that enacts the first agent's intention. Suppose that the first agent intends to raise his arm. Thanks to the second agent's intervention, his arm goes up. Does the first agent perform an intentional act? Indeed, does *he* raise his arm, or does the second agent raise it?

I think we need to consider two distinct scenarios:

(A) Suppose that the second agent is committed to sending the appropriate signals to the device that will guarantee the execution of the first agent's intention—suppose the second agent is a technician who is assigned the job of satisfying the intentions of the first agent without questioning. Under this description a case can be made for the claim that the first agent raised his arm intentionally. The case is perhaps clearest if we replace the second agent with a robot that is programmed to press the appropriate buttons when it identifies the intention of the first agent. This will be so even if the robot is prone to losing power and, as a result, sometimes becoming inactive (or if the committed second agent sometimes falls asleep on the job).

(B) Suppose that the second agent is capable of evaluating the intentions of the first agent and is empowered to censor the first agent when she judges the act to be impermissible. Now when the first agent intends to raise his arm and she sees nothing wrong with the act and presses the right buttons and the first agent's arm goes up, the first agent did not act intentionally. In fact, I am inclined to say that she (the second agent) raised his arm for him.

The difference between scenarios (A) and (B) can be captured by considering the two systems from the perspective of their respective functions. In (A) the second agent is a mere extension of the prosthetic device, and the way she acts merely replaces a circuit in the system. Her role is no different from the role of the second agent in cases B(1) and B(2). The system incorporating the two agents and the prosthetic device is functioning exactly the way the first agent plus the prosthetic device were supposed to, which explains why we judge the acts produced in scenario (A) to be intentional. But in scenario (B), the function of the whole system is no longer that of satisfying the intentions of the first agent. As I stipulated at the beginning, intentions are formed in causal consequence of a process of deliberation. The deliberative system takes in as inputs information concerning the circumstances, consults a series of instrumental beliefs, and a preference ordering of outcomes and chooses a course of action that involves the most preferred consequences. The system comprised of the first agent plus any prosthetic devices he might be plugged into can be said to have the function of satisfying whatever intention is reached through the process of deliberation by executing the basic act that is part of the content of the intention. Only when the system executes this function (in the way it is supposed to) is the action intentional. (Consider how when a demon out of the blue induces an intention in me—an intention that given my beliefs and desires I myself would not have formed—and the intention causes a bodily movement, I do not *act* intentionally.) When we couple this system to a second system that introduces a second deliberative process, we no longer retain the functional integrity of the first system.

As a result, when the first agent's arm goes up in response to his intention to raise his arm, this arm rising is the result of the intentions of the second agent too. This is a real case of "shared intentions"; both agents must have reached through a deliberative process an intention with the same content (i.e., the intention to raise the first agent's arm), and the basic act is actually "in response" to both intentions, which explains why it is false to say that it was an intentional act of the first agent. There was no *one* agent of whom it was an intentional act.

References

Audi, R. (1973). Intending. *Journal of Philosophy*, *70*, 387-403.

Bach, K. (1978). A representational theory of action. *Philosophical Studies*, *34*, 361-379.

Brand, M. (1984). *Intending and acting: Toward a naturalized action theory*. Cambridge, MA: MIT Press.

Bishop, J. (1989). *Natural agency*. Cambridge: Cambridge University Press.

Chisholm, R. M. (1966). Freedom and action. In K. Lehrer (Ed.), *Freedom and determinism* (pp. 11-44). New York: Random House.

Enç, B. (1979). Function attributions and functional explanation. *Philosophy of Science*, *46*, 343-365.

Enç, B. , & Adams, F. (1992). Functions and goal-directedness. *Philosophy of Science*, *59*, 635-654.

Davidson, D. (1980). Freedom to act. In *Essays on actions and events* (pp. 63-81). Oxford: Oxford University Press.

Davies, M. (1963). Function in perception. *Australasian Journal of Philosophy*, *61*, 409-426.

Dretske, F., & Snyder, A. (1972). Causal irregularity. *Philosophy of Science*, *39*, 69-71.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, *66*, 828-839.

Gibbons, J. (2001). Knowledge in action. *Philosophy and Phenomenological Research*, *62*, 579-600.

Ginet, C. (1990). *On action*. Cambridge: Cambridge University Press.

Mele, A. R. (1987). Intentional action and wayward causal chains: The problem of tertiary waywardness. *Philosophical Studies*, *51*, 55-60.

Mele, A. R. (1992). *Springs of action*. Oxford: Oxford University Press.

Mele, A. R., & Moser, R. K. (1994). Intentional action. *Noϕs*, *28*, 39-68.

Millikan, R. (1989). In defense of proper functions. *Philosophy of Science*, *56*, 288-302.

Moya, C. J. (1990). *The philosophy of action: An introduction*. Cambridge, England: Polity Press.

Peacocke, C. (1979). *Holistic explanation: Action, space, interpretation*. Oxford: Clarendon Press.

Pears, D. (1975). The appropriate causation of intentional basic actions. *Critica*, *7*, 39-69.

Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.

Taylor, R. (1966). *Action and purpose*. Englewood Cliffs, NJ: Prentice Hall.

Tuomela, R. (1977). *Human action and its explanation: A study on the philosophical foundations of psychology*. Dordrecht, Holland: Reidel.