THE CONCEIVABILITY OF MECHANISM

1. BY "mechanism" I am going to understand a special application of physical determinism—namely, to all organisms with neurological systems, including human beings. The version of mechanism I wish to study assumes a neurophysiological theory which is adequate to explain and predict all movements of human bodies except those caused by outside forces. The human body is assumed to be as complete a causal system as is a gasoline engine. Neurological states and processes are conceived to be correlated by general laws with the mechanisms that produce movements. Chemical and electrical changes in the nervous tissue of the body are assumed to cause muscle contractions, which in turn cause movements such as blinking, breathing, and puckering of the lips, as well as movements of fingers, limbs, and head. Such movements are sometimes produced by forces (pushes and pulls) applied externally to the body. If someone forced my arm up over my head, the theory could not explain that movement of my arm. But it could explain any movement not due to an external push or pull. It could explain, and predict, the movements that occur when a person signals a taxi, plays chess, writes an essay, or walks to the store.[1]

It is assumed that the neurophysiological system of the human body is subject to various kinds of stimulation. Changes of temperature or pressure in the environment; sounds, odors; the ingestion of foods and liquids: all these will have an effect on the nerve pulses that turn on the movement-producing mechanisms of the body.

2. The neurophysiological theory we are envisaging would, as said, be rich enough to provide systematic causal explanations of all bodily movements not due to external physical causes. These explanations should be understood as stating *sufficient* conditions

[1] If you said "Get up!" and I got up, the theory would explain my movements in terms of neurophysiological events produced by the impact of sound waves on my auditory organs.

of movement and not merely necessary conditions. They would employ laws that connect neurophysiological states or processes with movements. The laws would be universal propositions of the following form: whenever an organism of structure S is in state q it will emit movement m. Having ascertained that a given organism is of structure S and is in state q, one could deduce the occurrence of movement m.

It should be emphasized that this theory makes no provision for desires, aims, goals, purposes, motives, or intentions. In explaining such an occurrence as a man's walking across a room, it will be a matter of indifference to the theory whether the man's purpose, intention, or desire was to open a window, or even whether his walking across the room was intentional. This aspect of the theory can be indicated by saying that it is a "nonpurposive" system of explanation.

The viewpoint of mechanism thus assumes a theory that would provide systematic, complete, nonpurposive, causal explanations of all human movements not produced by external forces. Such a theory does not at present exist. But nowadays it is ever more widely held that in the not far distant future there will be such a theory—and that it will be true. I will raise the question of whether this is conceivable. The subject belongs to an age-old controversy. It would be unrealistic for me to hope to make any noteworthy contribution to its solution. But the problem itself is one of great human interest and worthy of repeated study.

3. To appreciate the significance of mechanism, one must be aware of the extent to which a comprehensive neurophysiological theory of human behavior would diverge from those everyday explanations of behavior with which all of us are familiar. These explanations refer to purposes, desires, goals, intentions. "He is running to catch the bus." "He is climbing the ladder in order to inspect the roof." "He is stopping at this store because he wants some cigars." Our daily discourse is filled with explanations of behavior in terms of the agent's purposes or intentions. The behavior is claimed to occur in order that some state of affairs should be brought about or avoided—that the bus should be caught, the roof inspected, cigars purchased. Let us say that these are "purposive" explanations.

We can note several differences between these common purposive explanations and the imagined neurophysiological explanations. First, the latter were conceived by us to be systematic—that is, to belong to a comprehensive theory—whereas the familiar purposive explanations are not organized into a theory. Second, the neurophysiological explanations do not employ the concept of purpose or intention. Third, the neurophysiological explanations embody contingent laws, but purposive explanations do not.

Let us dwell on this third point. A neurophysiological explanation of some behavior that has occurred is assumed to have the following form:

> Whenever an organism of structure *S* is in neurophysiological state *q* it will emit movement *m*.
> Organism *O* of structure *S* was in neurophysiological state *q*.
> Therefore, *O* emitted *m*.[2]

The general form of purposive explanation is the following:

> Whenever an organism *O* has goal *G* and believes that behavior *B* is required to bring about G, *O* will emit *B*.
> *O* had *G* and believed *B* was required of *G*.
> Therefore, *O* emitted *B*.

Let us compare the first premise of a neurophysiological explanation with the first premise of a purposive explanation. The first premise of a neurophysiological explanation is a contingent proposition, but the first premise of a purposive explanation is not a contingent proposition. This difference will appear more clearly if we consider how, in both cases, the first premise would have to be qualified in order to be actually true. In both cases a *ceteris paribus* clause must be added to the first premise, or at least be implicitly understood. (It will be more perspicuous to translate "*ceteris paribus*" as "provided there are no

[2] A neurophysiological *prediction* would be of the same form, with these differences: the second premise would say that *O* is or will be in state *q* (instead of *was*), and the conclusion would say that *O will* emit *m* (instead of *emitted*).

countervailing factors" rather than as "other things being equal.")

Let us consider what "*ceteris paribus*" will mean in concrete terms. Suppose a man climbed a ladder leading to a roof. An explanation is desired. The fact is that the wind blew his hat onto the roof and he wants it back. The explanation would be spelled out in detail as follows:

> If a man wants to retrieve his hat and believes this requires him to climb a ladder, he will do so provided there are no countervailing factors.
> This man wanted to retrieve his hat and believed that this required him to climb a ladder, and there were no countervailing factors.
> Therefore, he climbed a ladder.

What sorts of things might be included under "countervailing factors" in such a case? The unavailability of a ladder, the fear of climbing one, the belief that someone would remove the ladder while he was on the roof, and so on. (The man's failure to climb a ladder would *not* be a countervailing factor.)

An important point emerging here is that the addition of the *ceteris paribus* clause to the first premise turns this premise into an a priori proposition. If there were no countervailing factors whatever (if the man knew a ladder was available, had no fear of ladders or high places, no belief that he might be marooned on the roof, and so on); if there were no hindrances or hazards, real or imagined, physical or psychological; then if the man did not climb a ladder it would not be true that he *wanted* his hat back, or *intended* to get it back.[3]

[3] The correct diagnosis of such a failure will not be evident in all cases. Suppose a youth wants to be a trapeze performer in a circus, and he believes this requires daily exercise on the parallel bars. But he is lazy and frequently fails to exercise. Doesn't he really have the goal he professes to have: is it just talk? Or doesn't he really believe in the necessity of the daily exercise? Or is it that he has the goal and the belief and his laziness is a genuine countervailing factor? One might have to know him very well in order to give the right answer. In some cases there might be no definite right answer.

In his important recent book, *The Explanation of Behaviour*, Charles Taylor puts the point as follows:

> This is part of what we mean by "intending *X*," that, in the absence of interfering factors, it is followed by doing *X*. I could not be said to intend *X* if, even with no obstacles or other countervailing factors, I still didn't do it.[4]

This feature of the meaning of "intend" also holds true of "want," "purpose," and "goal."

Thus the universal premise of a purposive explanation is an a priori principle, not a contingent law. Some philosophers have made this a basis for saying that a purposive explanation is not a causal explanation.[5] But this is a stipulation (perhaps a useful one), rather than a description of how the word "cause" is actually used in ordinary language.

Let us consider the effect of adding a *ceteris paribus* clause to the universal premise of a neural explanation of behavior. Would a premise of this form be true a priori? Certainly not. Suppose it were believed that whenever a human being is in neural state *q* his right hand will move up above his head, provided there are no countervailing factors. What could be countervailing factors? That the subject's right arm is broken or that it is tied to his side, and so on. But the exclusion of such countervailing factors would have no tendency to make the premise true a priori. There is no connection of meaning, explicit or implicit, between the description of any neural state and the description of any movement of the hand. No matter how many countervailing factors are excluded, the proposition will not lose the character of a contingent law (unless, of course, we count the failure of the hand to move as itself a countervailing factor, in which case the premise becomes a tautology).

4. Making explicit the *ceteris paribus* conditions points up the different logical natures of the universal premises of the two kinds of explanation. Premises of the one sort express contingent corre-

[4] Charles Taylor, *The Explanation of Behaviour* (New York, 1964), p. 33.

[5] E.g., Taylor says that the agent's intention is not a "causal antecedent" of his behavior, for intention and behavior "are not contingently connected in the normal way" (*ibid.*).

lations between neurological processes and behavior. Premises of the other sort express a priori connections between intentions (purposes, desires, goals) and behavior.

This difference is of the utmost importance. Some students of behavior have believed that purposive explanations of behavior will be found to be less basic than the explanations that will arise from a future neurophysiological theory. They think that the principles of purposive explanation will turn out to be dependent on the neurophysiological laws. On this view our ordinary explanations of behavior will often be true: but the neural explanations will also be true—and they will be *more fundamental*. Thus we could, theoretically, *by-pass* explanations of behavior in terms of purpose, and the day might come when they simply fall into disuse.

I wish to show that neurophysiological laws could not be more basic than purposive principles. I shall understand the statement that a law L_2 is "more basic" than a law L_1 to mean that L_1 is dependent on L_2 but L_2 is not dependent on L_1. To give an example, let us suppose there is a uniform connection between food abstinence and hunger: that is, going without food for n hours always results in hunger. This is L_1. Another law L_2 is discovered—namely, a uniform connection between a certain chemical condition of body tissue (called "cell-starvation") and hunger. Whenever cell-starvation occurs, hunger results. It is also discovered that L_2 is more basic than L_1. This would amount to the following fact: food abstinence for n hours will not result in hunger *unless* cell-starvation occurs; and if the latter occurs, hunger will result *regardless of whether* food abstinence occurs. Thus the L_1 regularity is contingently dependent on the L_2 regularity, and the converse is not true. Our knowledge of this dependency would reveal to us the conditions under which the L_1 regularity would no longer hold.

Our comparison of the differing logical natures of purposive principles and neurophysiological laws enables us to see that the former cannot be dependent on the latter. The a priori connection between intention or purpose and behavior cannot fail to hold. It cannot be contingently dependent on any contingent regularity. The neurophysiological explanations of behavior could not, in the

sense explained, turn out to be more basic than our everyday purposive explanations.[6]

5. There is a second important consequence of the logical difference between neurophysiological laws and purposive principles. Someone might suppose that although purposive explanations cannot be dependent on nonpurposive explanations, they would be refuted by the verification of a comprehensive neurophysiological theory of behavior. I think this view is correct: but it is necessary to understand what it *cannot* mean. It cannot mean that the principles (the universal premises) of purposive explanations would be proved false. They cannot be proved false. It could not fail to be true that if a person wanted X and believed Y was necessary for X, and there were absolutely no countervailing factors, he would do Y.[7] This purposive principle is true a priori, not because of its form but because of its meaning—that is, because of the connection of meaning between the words "He wanted X and he realized that Y was necessary for X" and the words "He did Y." The purposive principle is not a law of nature but a conceptual truth. It cannot be confirmed or refuted by experience. Since the verification of a neurophysiological theory could never *disprove* any purposive principles, the only possible outcome of such verification, logically speaking, would be to prove that the purposive principles have no application to the world. I shall return to this point later.

6. We must come to closer grips with the exact logical relationship between neural and purposive explanations of behavior. Can explanations of both types be true of the same bit of behavior

---

[6] Taylor puts the point as follows :

> Because explanation by intentions or purposes is like explanation by an "antecedent" which is non-contingently linked with its consequent, i.e., because the fact that the behaviour follows from the intention other things being equal is not a contingent fact, we cannot account for this fact by more basic laws. For to explain a fact by more basic laws is to give the regularities on which this fact causally depends. But not being contingent, the dependence of behaviour on intention is not contingent on anything, and hence not on any such regularities [*ibid.*, p. 44].

[7] This is true if we use "wants X" to mean "is aiming at X." But sometimes we may mean no more than "would like to have X," which may represent a mere wish.

on one and the same occasion? Is there any rivalry between them? Some philosophers would say not. They would say that, for one thing, the two kinds of explanation explain different things. Purposive explanations explain actions. Neurophysiological explanations explain movements. Both explain behavior: but we can say this only because we use the latter word ambiguously to cover both actions and movements. For a second point, it may be held that the two kinds of explanation belong to different "bodies of discourse" or to different "language games." They employ different concepts and assumptions. One kind of explanation relates behavior to causal laws and to concepts of biochemistry and physiology, to nerve pulses and chemical reactions. The other kind of explanation relates behavior to the desires, intentions, goals, and reasons of persons. The two forms of explanation can co-exist, because they are irrelevant to one another.[8]

It is true that the two kinds of explanation employ different concepts and, in a sense, explain different things: but are they really independent of one another? Take the example of the man climbing a ladder in order to retrieve his hat from the roof. This explanation relates his climbing to his intention. A neurophysiological explanation of his climbing would say nothing about his intention but would connect his movements on the ladder with chemical changes in body tissue or with the firing of neurons. Do the two accounts interfere with one another?

7. I believe there *would* be a collision between the two accounts if they were offered as explanations of one and the same occurrence of a man's climbing a ladder. We will recall that the envisaged

[8] The following remarks by A. I. Melden present both of these points:

> Where we are concerned with causal explanations, with events of which the happenings in question are effects in accordance with some law of causality, to that extent we are not concerned with human actions at all but, at best, with bodily movements or happenings; and where we are concerned with explanations of human action, there causal factors and causal laws in the sense in which, for example, these terms are employed in the biological sciences are wholly irrelevant to the understanding we seek. The reason is simple, namely, the radically different logical characteristics of the two bodies of discourse we employ in these distinct cases—the different concepts which are applicable to these different orders of inquiry [A. I. Melden, *Free Action* (New York, 1961), p. 184].

neurophysiological theory was supposed to provide *sufficient* causal explanations of behavior. Thus the movements of the man on the ladder would be *completely* accounted for in terms of electrical, chemical, and mechanical processes in his body. This would surely imply that his desire or intention to retrieve his hat had nothing to do with his movement up the ladder. It would imply that on this same occasion he would have moved up the ladder in exactly this way even if he had had no intention to retrieve his hat, or even no intention to climb the ladder. To mention his intention or purpose would be no explanation, nor even part of an explanation, of his movements on the ladder. Given the antecedent neurological states of his bodily system together with general laws correlating those states with the contractions of muscles and movements of limbs, he would have moved as he did regardless of his desire or intention. If every movement of his was completely accounted for by his antecedent neurophysiological states (his "programming"), then it was not true that those movements occurred *because* he wanted or intended to get his hat.

8. I will briefly consider three possible objections to my claim that if mechanism were true the man would have moved up the ladder as he did even if he had not had any intention to climb the ladder. The first objection comes from a philosopher who espouses the currently popular psychophysical identity thesis. He holds that there is a neural condition that causes the man's movements up the ladder, and he further holds that the man's intention to climb the ladder (or, possibly, his having the intention) is contingently identical with the neural condition that causes the movements. Thus, if the man had not intended to climb the ladder, the cause of his movements would not have existed, and so those movements would not have occurred. My reply would be that the view that there may be a contingent identity (and not merely an extensional equivalence) between an intention (or the having of the intention) and a neural condition is not a meaningful hypothesis. One version of the identity thesis is that *A*'s intention to climb the ladder is contingently identical with some process in *A*'s brain. Verifying this identity would require the meaningless step of trying to discover whether *A*'s intention is located in his

brain. One could give meaning to the notion of the location of *A*'s intention in his brain by stipulating that it has the same location as does the correlated neural process. But the identity that arose from this stipulation would not be contingent.[9] Another version of the identity thesis is that the event of Smith's having the intention *I* is identical with the event of Smith's being in neural condition *N*. This version avoids the above "location problem": but it must take on the task (which seems hopeless) of explaining how the property "having intention *I*" and the property "being in neural condition *N*" could be contingently identical and not merely co-extensive.[10]

The second objection comes from an epiphenomenalist. He holds that the neurophysiological condition that contingently causes the behavior on the ladder also contingently causes the intention to climb the ladder, but that the intention stands in no causal relation to the behavior. If the intention had not existed, the cause of it and of the behavior would not have existed, and so the behavior would not have occurred. A decisive objection to epiphenomenalism is that, according to it, the relation between intention and behavior would be purely contingent. It would be conceivable that the neurophysiological condition that always causes ladder-climbing movements should also always cause the intention to *not* climb up a ladder. Epiphenomenalism would permit it to be universally true that whenever any person intended to *not* do any action, he did it, and that whenever any person intended to do any action, he did not do it. This is a conceptual absurdity.

The third objection springs from a philosopher who combines mechanism with logical behaviorism. He holds that some condition of the neurophysiological system causes the preparatory movements, gestures, and utterances that are expressions of the man's intention to climb the ladder; and it also causes his move-

[9] This point is argued in my "Scientific Materialism and The Identity Theory," *Dialogue*, 3 (1964); also in my forthcoming monograph, *Problems of Mind*, sec. 18., to be published in the Harper Guide to Philosophy, edited by Arthur Danto.

[10] For an exposition of this problem see Jaegwon Kim's "On the Psycho-Physical Identity Theory," *American Philosophical Quarterly*, 3 (1966).

ments up the ladder. The component of logical behaviorism in his over-all view is this: he holds that the man's having the intention to climb the ladder is simply a logical construction out of the occurrence of the expressions of intention and also the occurrence of the ladder-climbing movements. Having the intention is nothing other than the expressive behavior plus the subsequent climbing behavior. Having the intention is defined in terms of behavior-events that are contingently caused by a neurophysiological condition. The supposition that the man did not have the intention to climb the ladder would be identical with the supposition that either the expressive behavior or the climbing behavior, or both, did not occur. If either one did not occur, then neither occurred, since by hypothesis both of them have the same cause. Thus it would be false that the man would have moved up the ladder as he did even if he had not had an intention to climb the ladder.

I think that this third position gives an unsatisfactory account of the nature of intention. Actually climbing the ladder is not a necessary condition *simpliciter* for the existence of the intention to climb the ladder. It is a necessary condition *provided* there are no countervailing factors. But there is no definite number of countervailing factors, and so they cannot be exhaustively enumerated. In addition, some of them will themselves involve the concepts of desire, belief, or purpose. For example: a man intends to climb the ladder, but also he does not want to look ridiculous; as he is just about to start climbing he is struck by the thought that he will look ridiculous; so he does not climb the ladder, although he had intended to. An adequate logical behaviorism would have to analyze away not only the initial reference to intention, but also the reference to desire, belief, purpose, and all other psychological concepts, that would occur in the listing of possible countervailing factors. There is no reason for thinking that such a program of analysis could be carried out.

Thus a mechanist can hope to avoid the consequence that the man would have moved up the ladder as he did even if he had not had the intention of climbing the ladder, by combining his mechanist doctrine with the psychophysical identity thesis, or with epiphenomenalism, or with logical behaviorism. But these

supplementary positions are so objectionable or implausible that the mechanist is not really saved from the above consequence.

9. Let us remember that the postulated neurophysiological theory is comprehensive. It is assumed to provide complete causal explanations for all bodily movements that are not produced by external physical forces. It is a closed system in the sense that it does not admit, as antecedent conditions, anything other than neurophysiological states and processes. Desires and intentions have no place in it.

If the neurophysiological theory were true, then in no cases would desires, intentions, purposes be necessary conditions of any human movements. It would never be true that a man would *not* have moved as he did if he had *not* had such and such an intention. Nor would it ever be true that a certain movement of his was due to, or brought about by, or caused by his having a certain intention or purpose. Purposive explanations of human bodily movements would *never* be true. Desires and intentions would not be even potential causes of human movements in the actual world (as contrasted with some possible world in which the neurophysiological theory did not hold true).

It might be thought that there could be two different systems of causal explanations of human movements, a purposive system and a neurophysiological system. The antecedent conditions in the one system would be the desires and intentions of human beings; in the other they would be the neurophysiological states and processes of those same human beings. Each system would provide adequate causal explanations of the same movements.

Generally speaking, it is possible for there to be a plurality of simultaneous sufficient causal conditions of an event. But if we bear in mind the comprehensive aspect of the neurophysiological theory—that is, the fact that it provides sufficient causal conditions for all movements—we shall see that desires and intentions could not be causes of movements. It has often been noted that to say *B causes C* does not mean merely that whenever *B* occurs, *C* occurs. Causation also has subjunctive and counterfactual implications: if *B were to* occur, *C would* occur; and if *B* had *not* occurred, *C* would *not* have occurred. But the neurophysiological theory would provide sufficient causal conditions for every human

movement, and so there would be no cases at all in which a certain movement would not have occurred if the person had not had this desire or intention. Since the counterfactual would be false in all cases, desires and intentions would not be causes of human movements. They would not ever be sufficient causal conditions nor would they ever be necessary causal conditions.

10. Let us tackle this immensely important point from a different angle. Many descriptions of behavior ascribe actions to persons: they say that someone *did* something—for example, "He signed the check," "You lifted the table," "She broke the vase." Two things are implied by an ascription of an "action" to a person[11]: first, that a certain state of affairs came into existence (his signature's being present on the check, the table's being lifted, the vase's being broken); second, that the person intended that this state of affairs should occur. If subsequently we learn that not both conditions were satisfied, either we qualify the ascription of action or reject it entirely. If the mentioned state of affairs did not come into existence (for example, the vase was not broken), then the ascription of action ("She broke the vase") must be withdrawn. If it did come into existence but without the person's intention, then the ascription of action to the person must be diminished by some such qualification as "unintentionally" or "accidentally" or "by mistake" or "inadvertently," it being a matter of the circumstances which qualification is more appropriate. A qualified ascription of action still implies that the person played some part in bringing about the state of affairs—for example, her hand struck the vase. If she played no part at all, then it cannot rightly be said, even with qualification, that she broke the vase.

Suppose a man intends to open the door in front of him. He turns the knob and the door opens. Since turning the knob is what normally causes the door to open, we should think it right to say that *he* opened the door. Then we learn that there is an electric mechanism concealed in the door which caused the door to open at the moment he turned the knob, and furthermore that

[11] I am following Charles Taylor here: *op. cit.*, pp. 27-32.

there is no causal connection between the turning of the knob and the operation of the mechanism. So his act of turning the knob had nothing to do with the opening of the door. We can no longer say that *he* opened the door: nothing he did had any causal influence on that result. We might put the matter in this way: because of the operation of the electric mechanism he had no opportunity to open the door.

The man of our example could say that at least he turned the knob. He would have to surrender this claim, however, if it came to light that still another electrical mechanism caused the knob to turn when it did, independently of the motion of his hand. The man could assert that, in any case, he moved his hand. But now the neurophysiological theory enters the scene, providing a complete causal explanation of the motion of his hand, without regard to his intention.

The problem of what to say becomes acute. Should we deny that he moved his hand? Should we admit that he moved his hand, but with some qualification? Or should we say, without qualification, that he moved his hand?

11. There is an important similarity between our three examples and an important difference. The similarity is that in all three cases a mechanism produced the intended states of affairs, and nothing the agent did had any influence on the operation of the mechanism. But there is a difference between the cases. In each of the first two, we can specify something the man did (an action) which would normally cause the intended result to occur, but which did not have that effect on this occasion. The action in the first case was turning the knob, and in the second it was gripping the knob and making a turning motion of the hand. In each of these cases there was an action, the causal efficacy of which was nullified by the operation of a mechanism. Consequently, we can rightly say that the man's action *failed* to make a contribution to the intended occurrence, and so we can deny that *he* opened the door or turned the knob.

In the third case is there something the man did which normally causes that movement of the hand? What was it? When I move my hand in the normal way is there something else *I do* that causes my hand to move? No. Various events take place in my body

(for example, nerve pulses) but they cannot be said to be *actions* of mine. They are not things I do.

But in this third case the man *intended* to make a turning motion of his hand. Is this a basis for a similarity between the third case and the first two? Can we say that one's intention to move one's hand is normally a cause of the motion of one's hand, but that in our third case the causal efficacy of the intention was nullified by the operation of the neurophysiological mechanism?

On the question of whether intentions are causes of actions, Taylor says something that is both interesting and puzzling. He declares that to call something an action, in an unqualified sense "means not just that the man who displayed this behaviour had framed the relevant intention or had this purpose, but also that *his intending it brought it about.*"[12] Now to say that *A* "brings about" *B* is to use the language of causation. "Brings about" is indeed a synonym for "causes."

12. Is there any sense at all in which a man's intention to do something can be a cause of his doing it? In dealing with this point I shall use the word "cause" in its widest sense, according to which anything that explains, or partly explains, the occurrence of some behavioi is the cause, or part of the cause, of the behavior. To learn that a man intended to climb a ladder would not, in many cases, explain why he climbed it. It would not explain what he climbed it for, what his reason or purpose was in climbing it, whereas to say what his purpose was would, in our broad sense, give the cause or part of the cause of his climbing it.

In considering intention as a cause of behavior *X*, it is important to distinguish between the intention to do *X* (let us call this *simple intention*) and in the intention to do something else *Y* in or by doing *X* (let us call this *further intention*). To say that a man intended to climb a ladder would not usually give a cause of his climbing it; but stating his purpose in climbing it would usually be

[12] Taylor, *op. cit.*, p. 33 (my italics). Taylor says that an intention is *not* "a causal antecedent" of the intended behavior, for the reason that the intention and the behavior are not *contingently connected.* I think he may be fairly represented as holding that an intention does *cause* the intended behavior, although not in the sense of "cause" in which cause and effect are contingently correlated.

giving the (or a) cause of the action. It is a natural use of language to ask, "What caused you to climb the ladder?"; and it is an appropriate answer to say, "I wanted to get my hat." (*Question:* "Good heavens, what caused you to vote a straight Republican ticket?" *Answer:* "I wanted to restore the two-party system.") Our use of the language of causation is not restricted to the cases in which cause and effect are assumed to be contingently related.

13. Can the simple intention to do *X* ever be a cause of the doing of *X*? Can it ever be said that a person's intention to climb a ladder caused him to climb it, or brought about his action of climbing it? It is certainly true that whether a man does or does not intend to do *X* will make a difference in whether he will do *X*. This fact comes out strongly if we are concerned to predict whether he will do *X;* obviously, it would be important to find out whether he intends to do it. Does not this imply that his intention has "an effect on his behavior"?[13]

Commonly, we think of dispositions as causes of behavior. If with the same provocation one man loses his temper and another does not, this difference in their reactions might be explained by the fact that the one man, but not the other, is of an irritable disposition. If dispositions are causes, we can hardly deny the same role to intentions. Both are useful in predicting behavior. If I am trying to estimate the likelihood that this man is going to do so-and-so, the information that he has a disposition to do it in circumstances like these will be an affirmative consideration. I am entitled to give equal or possibly greater weight to the information that he intends to do it.

Not only do simple intentions have weight in predicting actions, but also they figure in the explanation of actions that have already occurred. If a man who has just been released from prison promptly climbs a flagpole, I may want an explanation of that occurrence. If I learn that he had previously made up his

[13] Taylor's phrase, *op. cit.*, p. 34. In my review of Taylor's book ("Explaining Behavior," *Philosophical Review*, LXXVI [1967], 97-104), I say that Taylor is wrong in holding that a simple intention *brings about* the corresponding behavior. But now I am holding that he is partly right and partly wrong: right about previously formed simple intentions, wrong about merely concurrent simple intentions.

mind to do it, but had been prevented by his imprisonment, I have received a partial explanation of why he is climbing the flagpole, even if I do not yet know his further intention, if any, in climbing it. In general, if I am surprised at an action, it will help me to understand its occurrence if I find out that the agent had previously decided to do it but was prevented by an obstacle which has just been removed.

14. The simple intentions so far considered were formed in advance of the corresponding action. But many simple intentions are not formed in advance of the corresponding action. Driving a car, one suddenly (and intentionally) presses the brake pedal: but there was no time before this action occurred when one intended to do it. The intention existed only at the time of the action, or only *in* the action. Let us call this a merely concurrent simple intention. Can an intention of this kind be a causal factor in the corresponding action?

Here we have to remember that if the driver did not press the brake intentionally, his pressing of the brake was not unqualified action. The presence of simple intention in the action (that is, its being intentional) is an analytically necessary condition for its being unqualified action. This condition is not a cause but a defining condition of unqualified action. If this condition were not fulfilled, one would have to use some mitigating phrase—for example, that the driver pressed the brake by mistake. Thus, a simple intention that is merely concurrent cannot be a cause of the corresponding action.

15. Can we not avoid committing ourselves to the assumption that the pressing of the driver's foot on the brake was either intentional or not intentional? Can we not think of it, in a neutral way, as merely behavior? Yes, we can. But it *was* either intentional or not intentional. If the latter, then there was no simple intention to figure as a cause of the behavior. If the former, then the behavior was action, and the driver's merely concurrent simple intention was a defining condition and not a cause of the behavior. The "neutral way" of thinking about the behavior would be merely incomplete. It would be owing to ignorance and not to the existence of a third alternative. It is impossible, by the definition of "action," that the behavior of pressing the brake should be an

action and yet not be intentional. Thus it is impossible that a merely concurrent simple intention should have caused the behavior of pressing the brake, whether the behavior was or was not action.

To summarize this discussion of intentions as causes: we need to distinguish between simple intentions and further intentions. If an agent does X with the further intention Y, then it is proper to speak of this further intention as the (or a) cause of the doing of X. Simple intentions may be divided into those that are formed prior to the corresponding actions, and those that are merely concurrent with the actions. By virtue of being previously formed, a simple intention can be a cause of action. But in so far as it is merely concurrent, a simple intention cannot be a cause of the corresponding action.

16. Let us try now to appraise Taylor's view as to the causal role of intention in behavior. He holds that it would not be true, without qualification, that one person stabbed another unless his intention to stab him "brought about" the stabbing (*ibid.*, p. 33). The example was meant to be of a previously formed intention—for Taylor speaks of the agent's *deciding* to stab someone. But a majority of actions do not embody intentions formed in advance. They embody merely concurrent intentions. The latter cannot be said to cause (bring about) the corresponding actions. Possibly because he has fixed his attention too narrowly on cases of decision, Taylor errs in holding that, in general, the concept of action requires that the agent's intention should have brought about the behavior. When the action is merely intentional (without previous intention) the agent's intention cannot be said to bring about his behavior. In such cases his intention gives his behavior the character of *action*, but it does this by virtue of being a defining condition of action, not by virtue of being a cause of either behavior or action.

17. Our reflections on the relationship of intention to behavior arose from a consideration of three examples of supposed action—opening a door, turning a knob, making a turning motion of the hand. In the first two cases we imagined mechanisms that produced the intended results independently of the agent's intervention. Consequently, we had to deny that *he* opened the door

or turned the knob. Then we imagined a neurophysiological cause of the motion of his hand, and we asked whether this would imply, in turn, that *he* did not move his hand.

Is the movement of his hand independent of his "intervention" by virtue of being independent of his intention? We saw previously (Section 8) that a comprehensive neurophysiological theory would leave no room for desires and intentions as causal factors. Consequently, neither the man's previously formed simple intention to move his hand nor his further intention (to open the door) could be causes of the movement of his hand.

18. We noticed before that it is true a priori that if a man wants Y, or has Y as a goal, and believes that X is required for Y, then in the absence of countervailing factors he will do X. It is also true a priori that if a man forms the intention (for example, decides) to do X, then in the absence of countervailing factors he will do X. These a priori principles of action are assumed in our everyday explanations of behavior.

We saw that mechanistic explanations could not be more basic than are explanations in terms of intentions or purposes.

We saw that the verification of mechanistic laws could not disprove the a priori principles of action.

Yet a mechanistic explanation of behavior rules out any explanation of it in terms of the agent's intentions. If a comprehensive neurophysiological theory is true, then people's intentions never are causal factors in behavior.

19. Thus if mechanism is true, the a priori principles of action do not apply to the world. This would have to mean one or the other of two alternatives. The first would be that people do not have intentions, purposes, or desires, or that they do not have beliefs as to what behavior is required for the fulfillment of their desires and purposes. The second alternative would be that although they have intentions, beliefs, and so forth, there always are countervailing factors—that is, factors that interfere with the operation of intentions, desires, and decisions.

The second alternative cannot be taken seriously. If a man wants to be on the opposite bank of a river and believes that swimming it is the only thing that will get him there, he will swim it unless there are countervailing factors, such as an inability

to swim or a fear of drowning or a strong dislike of getting wet. In this sense it is not true that countervailing factors are present *whenever* someone has a goal. There are not *always* obstacles to the fulfillment of any purpose or desire.

It might be objected that mechanistic causation itself is a universal countervailing factor. Now if this were so it would imply that purposes, intentions, and desires never have any effect on behavior. But it is not a coherent position to hold that some creatures have purposes and so forth, yet that these have no effect on their behavior. Purposes and intentions are, in concept, so closely tied to behavioral effects that the total absence of behavioral effects would mean the total absence of purposes and intentions. Thus the only position open to the exponent of mechanism is the first alternative—namely, that people do not have intentions, purposes, or beliefs.

What I have called "a principle of action" is a conditional proposition, having an antecedent and a consequent. The whole conditional is true a priori, and therefore if the antecedent holds in a particular case, the consequent must also hold in that case. To say that the antecedent holds in a particular case means that it is true of some person (or animal). It means that the person has some desire or intention, and also has the requisite belief. If this were so, and if there were no countervailing factors, it would follow that the person would act in an appropriate manner. His intention or desire would, in our broad sense, be a cause of his action—that is, it would be a factor in the explanation of the occurrence of the action.

But this is incompatible with mechanism. A mechanist must hold, therefore, that the principles of action have no application to reality, in the sense that no one has intentions or desires or beliefs.

Some philosophers would regard this result as an adequate refutation of mechanism. But others would not. They would say that the confirmation of a comprehensive neurophysiological theory of behavior is a logical possibility, and therefore it is logically possible that there are no desires, intentions, and so forth, and that to deny these logical possibilities is to be dogmatic and antiscientific. I will avoid adopting this "dogmatic" and

"antiscientific" position, and will formulate a criticism of mechanism from a more "internal" point of view.

20. I wish to make a closer approach to the question of the conceivability of mechanism. We have seen that mechanism is incompatible with purposive behavior, but we have not yet established that it is incompatible with the existence of merely intentional behavior. A man can do something intentionally but with no further intention: his behavior is intentional but not purposive. One possibility is that this behavior should embody a merely concurrent simple intention. Since such intentions are not causes of the behavior to which they belong, their existence does not appear to conflict with mechanistic causation. Mechanism's incompatibility with purposive behavior has not yet shown it to be incompatible with intentional behavior as such.

But could it be true that sometimes people acted intentionally although it was never true that they acted for any purpose? Could they do things intentionally but never with any further intention?

If some intentional actions are purposeless, it does not follow that all of them could be purposeless. And I do not think this is really a possibility. I will not attempt to deal with every kind of action. But consider that subclass of actions that are activities. Any physical activity is analyzable into components. If a man is painting a wall, he is grasping a brush, dipping the brush into the paint, moving his arm back and forth. He does these things in painting. They are parts of his activity of painting. If someone is rocking in a chair, he is pushing against the floor with his feet, and pressing his back against the back of the chair. These are subordinate activities in the activity of rocking. If the one who is painting is asked why he is dipping the brush into the paint, he can answer, "I am painting this wall." This is an explanation of what he is doing in dipping the brush, and also of what he is dipping it *for*. It is a purposive explanation. A person can put paint on a wall, or rock in a chair, or pace back and forth, without having any purpose in doing so. Still these activities could be intentional, although not for any purpose.

Whether intentional or not, these activities would be analyzable into component parts. If the activity is intentional, then at least

some of its components will be intentional. If none of them were, the whole to which they belong would not be intentional. A man could not be intentionally putting paint on a wall if he did not intentionally have hold of a brush. Now this is not strictly true since he might not be aware that he was holding a *brush*, rather than a roller or a cloth. But there will have to be *some* description of what he is holding according to which it is true that he is intentionally holding it and intentionally dipping it in the paint.

Thus an intentional activity must have intentional components. The components will be purposive in relation to the whole activity. If X is an intentional component of Y, one can say with equal truth that in X-ing one is Y-ing, or that one is X-ing in order to Y. In moving the pencil on the paper one is drawing a figure: but also one is moving the pencil in order to draw a figure.

I conclude that if there could be no purposive behavior, there could be no intentional activities. Strictly speaking, this does not prove that there could be no intentional action, since many actions are not activities (for example, catching a ball or winning a race, as contrasted with playing ball or running in a race). But many of the actions that are not activities are stages in, or terminations of, activities and could not exist if the activities did not. Although I do not know how to prove the point for all cases, it seems to me highly plausible that if there could be no intentional activities there could be no intentional behavior of any sort—so plausible that I will assume it to be so. A life that was totally devoid of activities certainly could not be a human life. My conclusion is that since mechanism is incompatible with purposive behavior, it is incompatible with intentional activities, and consequently is incompatible with *all* intentional behavior.

21. The long-deferred question of whether the man of our example moved his hand on the doorknob will be answered as follows. The action of moving his hand cannot be rightly ascribed to him. It should not even be ascribed to him with some qualification such as "unintentionally" or "accidentally," for the use of these qualifications implies that there are cases in which it is right to say of a man that he did something "intentionally" or "purposely." But mechanism rules this out. On the other hand, to say "He did not move his hand" would be misleading, not only for the reason

just stated, but also for the further reason that this statement would normally carry the implication that his hand did not move—which is false. Neither the sentence "He moved his hand" nor the sentence "He did not move his hand" would be appropriate. We would, of course, say "He moved his hand" if we understood this as merely equivalent to "His hand moved." (It is interesting that we do use these two sentences interchangeably when we are observing someone whom we know to be asleep or unconscious: we are equally ready to say either "He moved his hand" or "His hand moved.") But if we came to believe in mechanism we should, in consistency, give up the ascribing of action, even in a qualified way.

22. We can now proceed directly to the question of whether mechanism is conceivable. Sometimes when philosophers ask whether a proposition is conceivable, they mean to be asking whether it is self-contradictory. Nothing in our examination has indicated that mechanism is a self-contradictory theory, and I am sure it is not. Logically speaking, the earth and the whole universe might have been inhabited solely by organisms of such a nature that all of their movements could have been completely explained in terms of the neurophysiological theory we have envisaged. We can conceive that the world might have been such that mechanism was true. In this sense mechanism is conceivable.

But there is a respect in which mechanism is not conceivable. This is a consequence of the fact that mechanism is incompatible with the existence of any intentional behavior. The speech of human beings is, for the most part, intentional behavior. In particular, stating, asserting, or saying that so-and-so is true requires the intentional uttering of some sentence. If mechanism is true, therefore, no one can state or assert anything. In a sense, no one can *say* anything. Specifically, no one can assert or state that mechanism is true. If anyone were to assert this, the occurrence of his intentional "speech act" would imply that mechanism is false.

Thus there is a logical absurdity in asserting that mechanism is true. It is not that the doctrine of mechanism is self-contradictory. The absurdity lies in the human act of asserting the doctrine. The occurrence of this act of assertion is inconsistent with the content

of the assertion. The mere proposition that mechanism is true is not self-contradictory. But the conjunctive proposition, "Mechanism is true and someone asserts it to be true," *is* self-contradictory. Thus anyone's assertion that mechanism is true is necessarily false. The assertion implies its own falsity by virtue of providing a counterexample to what is asserted.

23. A proponent of mechanism might claim that since the absurdity we have been describing is a mere "pragmatic paradox" and not a self-contradiction in the doctrine of mechanism, it does not provide a sense in which mechanism is inconceivable. He may say that the paradox is similar to the paradox of a man's asserting that he himself is unconscious. There is an inconsistency between this man's act of stating he is unconscious and what he states. His act of stating it implies that what he states is false. But this paradox does not establish that a man cannot be unconscious, or that we cannot conceive that a man should be unconscious.

Now there is some similarity between the paradox of stating that oneself is unconscious and the paradox of stating that mechanism is true. But there is an important difference. *I* cannot state, without absurdity, that *I* am unconscious. But anyone else can, without absurdity, state that I am unconscious. There is only one person (myself) whose act of stating this proposition is inconsistent with the proposition. But an assertion of mechanism by any person whomsoever is inconsistent with mechanism. That I am unconscious is not (in consistency) statable by me. The unstatability is relative to only one person. But the unstatability of mechanism is absolute.

Furthermore, no one can consistently assert that although mechanism is unstatable it may be true. For this assertion, too, would require an intentional utterance (speech act) and so would be incompatible with mechanism.

We have elucidated a sense in which mechanism can properly be said to be inconceivable. The sense is that no one can consistently assert (or state, or say) that mechanism is, or may be, true.

If someone were to insist on asserting that mechanism is or may be true, his only recourse (if he were to be consistent) would be to adopt a form of solipsism. He could claim that mechanism is true for other organisms but not for himself. In this way he would free

his assertion of inconsistency, but at the cost of accepting the embarrassments and logical difficulties of solipsism. He would also be repudiating the scientific respectability of mechanism by denying the universality of the envisaged neurophysiological laws.

24. Our criticism that mechanism is not a consistently statable doctrine is, of course, purely logical in nature. It consists in deducing a consequence of mechanism. Now one may feel that this consequence cannot refute mechanism or jeopardize its status as a scientific theory. It would seem to be up to science alone to determine whether or not there is a comprehensive neurophysiological theory to explain all bodily movements in accordance with universal laws. If scientific investigation should confirm such a theory, then so be it! To confirm it would be to confirm its consequences. If confirming the theory were to prove that people do not have desires, purposes, or goals, then this result would have to be swallowed, no matter how upsetting it would be not only to our ordinary beliefs but also to our ordinary concepts.

Almost anyone will feel some persuasiveness in this viewpoint. Determinism is a painful problem because it creates a severe tension between two viewpoints, each of which is strongly attractive: one is that the concepts of purpose, intention, and desire, of our ordinary language, cannot be rendered void by scientific advance; the other is that those concepts cannot prescribe limits to what it is possible for empirical science to achieve.

Let us see what would be the effect on our thinking of a scientific confirmation of mechanism. Suppose I am playing catch with a small boy. The ball escapes his grasp and he runs after it. Any observer would agree that the boy is running after the ball. This description implies that the purpose of the boy's running is to get the ball, or that he is running because he wants to capture the ball.

Now suppose a neurological technician could explain and predict every movement of the boy's limbs without regard to the whereabouts of the ball, solely in terms of the changing states of the boy's neurophysiological system. Or, what is worse, suppose the technician could control the boy's movements by altering the states of his central nervous system at will—that is, by "pro-

gramming." We can imagine that it should be impossible for us to tell in a given instance, by observation of the boy's outward behavior and circumstances, whether the boy's limbs were responding to programming or whether he was running in order to retrieve the ball. And suppose that in many instances when we thought the behavior was intentional, it was subsequently proved to us that exactly the same inner physiological processes occurred as on those occasions when the technician controlled the boy's movements. We can also suppose that the neurologist's predictions of behavior would be both more reliable and more accurate than are the predictions based on purposive assumptions.

If such demonstrations occurred on a massive scale, we should be learning that the principles of purposive explanation have a far narrower application than we had thought. On more and more occasions we (that is, each one of us) would be forced to regard other human beings as mechanisms. The ultimate outcome of this development would be that we should cease to think of the behavior of others as being influenced by desires and intentions.

25. Having become believers in mechanistic explanations of the behavior of others, could each of us also come to believe that mechanistic causation is the true doctrine for his own case? Not if we realized what this would imply, for each of us would see that he could not include himself within the scope of the doctrine. Saying or doing something *for a reason* (in the sense of grounds as well as in the sense of purpose) implies that the saying or doing is intentional. Since mechanism is incompatible with the intentionality of behavior, my acceptance of mechanism as true for myself would imply that I am incapable of saying or doing anything for a reason. There could be *a* reason (that is, a cause) but there could not be such a thing as *my* reason. There could not, for example, be such a thing as my reason for stating that mechanism is true. Thus my assertion of mechanism would involve a second paradox. Not only would the assertion be inconsistent, in the sense previously explained, but also it would imply that I am incapable of having rational grounds for asserting anything, including mechanism.

Once again we see that mechanism engenders a form of solipsism. In asserting mechanism I must deny its application to

my own case: for otherwise my assertion would imply that I could not be asserting mechanism on rational grounds.

26. Some philosophers hold that if mechanism is true then a radical revision of our concepts is required. We need to junk all such terms as "intentionally," "unintentionally," "purposely," "by mistake," "deliberately," "accidentally," and so on. The classifying of utterances such as "asserting," "repeating," "quoting," "mimicking," "translating," and so forth, would have to be abandoned. We should need an entirely new repertoire of descriptions of a sort that would be compatible with the viewpoint of mechanism.

I think these philosophers have not grasped the full severity of the predicament. If mechanism is true, not only should we give up speaking of "asserting," but also of "describing" or even of "speaking." It would not even be right to say that a person *meant* something by the noise that came from him. No marks or sounds would mean anything. There could not be *language.*

A proponent of mechanism should not think that at present we are using the wrong concepts and that a revision is called for. If he is right, we do not use concepts at all. There is nothing to revise—and nothing to say. The motto of a mechanist ought to be: One cannot speak, therefore one must be silent.

27. To conclude: We have uncovered two respects in which mechanism is not a conceivable doctrine. The first is that the occurrence of an act of asserting mechanism is inconsistent with mechanism's being true. The second is that the asserting of mechanism implies that the one who makes the assertion cannot be making it on rational grounds.

In order to avoid these paradoxes, one must deny that mechanism is universally true. One can hold that it is true for others but not for oneself. It is highly ironical that the affirmation of mechanism requires one to affirm its metaphysical and methodological opposite—solipsism.

The inconceivability of mechanism, in the two respects we have elucidated, does not establish that mechanism is false. It would seem, logically speaking, that a comprehensive neurophysiological theory of human behavior ought to be confirmable by scientific investigation. Yet the assertion that this confirmation

had been achieved would involve the two paradoxes we have elucidated. Mechanism thus presents a harsh, and perhaps insoluble, antinomy to human thought.

Concluding unscientific postscript: I must confess that I am not entirely convinced of the correctness of the position I have taken in respect of the crux of this paper—namely, the problem of whether it is possible for there to be both a complete neurophysiological explanation and also a complete purposive explanation of one and the same sequence of movements. I do not believe I have really proved this to be impossible. On the other hand, it is true that for me (and for others, too) a sequence of sounds tends to lose the aspect of speech (language) when we conceive of those sounds as being caused neurophysiologically (especially if we imagine a technician to be controlling the production of the sounds). Likewise, a sequence of movements loses the aspect of action. Is this tendency due to some false picture or to some misleading analogy? Possibly so; but also possibly not. Perhaps the publication of the present paper will be justified if it provokes a truly convincing defense of the compatibility of the two forms of explanation.[14]

NORMAN MALCOLM

Cornell University

[14] A number of people have read various versions of this paper and I have profited from their criticisms. I am especially indebted to Elizabeth Anscombe, Keith Donnellan, Philippa Foot, G. H. von Wright, and Ann Wilbur. They are not responsible for the mistakes I have retained.