A formal approach to the problem of free will and determinism

by

PETER VAN INWAGEN
(Syracuse University)

In this paper I will present simple formal statements of the theses of free will and universal causal determinism, and show that while these theses are not formal contraries or contradictories, there is nevertheless an important sense in which they are incompatible. It is, of course, not quite realistic to talk about *the* theses of free will and determinism, since philosophers have given many different senses to these terms. I shall therefore make only this claim for the formal notions of free will and determinism set forth below: they are sufficiently like what is often meant by "free will" and "determinism" in informal philosophical disputation that the question of their compatibility is philosophically interesting.

I

I shall begin by offering informal statements of the theses that the formal statements are intended to embody:

> To say that we have *free will* is to say that the future presents us with real alternatives. Very often, if not always, when a man must choose between A and B (e.g., between falsifying records in an attempt to deceive a superior who rightly suspects him of embezzling funds, and telling all), each alternative is *open* to him: he *can* act either way.
>
> *Determinism* is the thesis that if time could be "rolled back" to any past instant, and then allowed to "go forward again," then there is no question but what history would "repeat" itself: we could be certain that things would happen "again" just as they happened

> the "first time." For example, if God were to cause the world to revert to precisely its condition at the moment Harold's eye was pierced by a Norman arrow, and then leave the world once more to its own devices, then nine hundred six years later, I (or perhaps only someone indistinguishable from me?) should sit at this desk (or at its twin?) writing these same words.

What I shall *not* do is to try to translate these informal statements into some sort of symbolism. Their pictorial content is too rich and their cognitive content too spare and too confused for this to be possible. It is, rather, my hope that the formal statements of free will and determinism that follow will "satisfy" a person who would accept the preceding two paragraphs as articulations, successful insofar as they have content, of what he means by "free will" and "determinism." The formal notions will be satisfactory to a person who might express his ideas of free will and determinism as above if he feels that they provide him with a *replacement* for these ideas—if he feels that by doing his thinking about free will and determinism in the terms provided by the formal notions he has lost nothing of cognitive value (though perhaps something of pictorial or poetic value) and has gained something in the way of clarity and precision.

This is not to say that our formal statements will be as clear as anyone could wish. The notions behind the predicates that appear in the formal statements will be no more than roughed out, and that in the most informal and general way. But this is a virtue as well as a defect. If we were to be more precise in our specification of the relations expressed by the predicates we shall introduce, we should have to choose sides in dubious battle: we should have to answer such questions as, How are possible worlds to be identified and individuated?; What is a law of nature?; How are we to understand the concepts of agency and ability? Instead of trying to answer these questions, I shall simply assume that they *have* acceptable answers, and keep my remarks general enough to accommodate any consistent combination of answers to them.

Our formal statements will be constructed from three two-place predicates and one name. These, together with suggested English readings, are:

Nxy	*x* is nomologically congruent to *y*
Sxy	*x* shares a slice with *y*
Hxy	*x* has access to *y*
A	the actual world

In addition, we introduce by definition a one-place predicate "*D*" read "is deterministic":

$$Dx =_{\mathrm{df}} \exists y(Nyx) \;\&\; \forall y(Nyx \;\&\; Syx \,.\!\supset y = x).$$

The range of our variables includes all possible worlds, but does not include mere *possibilia,* that is, individual possible but non-actual things.

The name "*A*" denotes, of course, the actual world, the world of fact and not of counterfact, fiction, or myth, the world comprising those and only those states of affairs that obtain *in re* and not *in solo intellectu.*

The predicate "*S*" will represent the dyadic relation that holds between *x* and *y* if and only if *x* and *y* are possible worlds that are indistinguishable at at least one instant of time. *S* is symmetrical and reflexive, but non-transitive.[1]

We may think of *S* in the following way. Let us imagine a Leibnizian God, who somehow "stands outside" all possible worlds and is able somehow to "examine" them individually *sub specie aeternitatis.* Presumably, such a God would be able to restrict His examination of a world to (focus on, as it were) the way that world is at a single instant of time. If we find this way of speaking intelligible, then we may say that *S* holds between *x* and *y* if and only if *x* and *y* are possible worlds and there is some instant *t*

[1] If there are any nontemporal possible worlds—worlds in which there is no such thing as the passage of time—then, by stipulation, if either *x* or *y* is nontemporal, then *x* bears *S* to *y* if and only if *x* and *y* are identical. This stipulation has the result that all nontemporal worlds are deterministic, a result that I find intuitively satisfying. The following discussion of *S* will assume, for the sake of simplicity, that all worlds are temporal.

such that if God were to "examine *x* as it is at *t*" and "examine *y* as it is at *t*," He could observe, on the basis of these examinations alone, no difference between *x* and *y*.

Or, if we are willing to think of a (temporal) possible world as a dense sequence of three-dimensional instantaneous "slices," then we may say that *S* holds between *x* and *y* just in the case that they are possible worlds that have at least one slice in common—hence the suggested English reading of "*S*".[2]

We shall understand the predicate "*N*" to represent an equivalence relation that holds between *x* and *y* if and only if *x* and *y* are possible worlds in which "the laws of nature" are the same. An alternative reading of "*Nxy*" is: "what is physically necessary and impossible in *x* is what is physically necessary and impossible in *y*."[3] Examples of worlds that (given the truth of our present beliefs) do not bear *N* to the actual world are: worlds in which moving material objects sometimes undergo perfectly sharp 90° changes in direction; worlds in which information is sometimes transmitted faster than the speed of light in a vacuum; worlds in which the energy of a photon is not proportionate to its wavelength; worlds in which the speed of light, the charge on the electron, and the universal gravitational constant have grossly different values from the values we find in our physics texts. Of course, the notions of natural law and physical impossibility are very cloudy. I think that no one has succeeded in making these notions clear, and perhaps no one ever will; perhaps they are

[2] We may also interpret "*Sxy*" as "*x* and *y* are indistinguishable over some finite interval" or "there is a period in which the course of history in *x* exactly parallels the course of history in *y*". If we adopt this stronger sense for "*S*", we shall obtain a weaker thesis of determinism. The argument of this paper does not depend on whether the stronger or the weaker sense is given to "determinism".

[3] These modal terms must be understood in an *absolute* or *intrinsic*, rather than a *relative* sense. Thus, while it may be physically impossible *relative to* past or present circumstances that a certain falling body should not strike the ground, it is *absolutely* physically impossible (we presently suppose) that that body should move faster than light, or stop dead without transferring its momentum to other bodies. For a more careful and detailed statement of this distinction, see Wilfrid Sellars, "Fatalism and determinism," in K. Lehrer, ed., *Freedom and determinism* (New York: Random House, 1966), p. 163.

ultimately incoherent. If that is the case, however, then the thesis of determinism is incoherent. And, of course, if determinism is incoherent, then there is no problem of free will and determinism. I shall simply assume that at least one clear meaning can be given to the phrase "laws of nature" that is not utterly at variance with our preanalytic expectations about what sorts of propositions should (if true) be called laws of nature, and which, moreover, is definite enough to yield (in principle) yes-no answers to questions of the form, "Are the laws of nature the same in possible worlds *x* and *y*?" when sufficient information about the worlds in question is known.

Let us now examine the predicate "*D*". This predicate is intended to represent a property of some possible worlds (called "being deterministic") which may be informally characterized as follows: a world *x* is deterministic if and only if *x* itself is the *only* world that both shares a slice with *x* and is nomologically congruent to *x*. Let us look at an example. Let W_1 be some possible world that shares with the actual world *A* a slice taken at the instant Harold's eye was pierced by a Norman arrow. W_1 may share indenumerably many other slices with *A*; it shares at least that slice. And let us suppose that in *A* and W_1 the laws of nature are the same. There are two possibilities: W_1 may *be* *A*, or it may be some other possible world. I shall try to indicate why I find it intuitively plausible to call W_1 and *A* "deterministic" only if they are identical.

Suppose W_1 and *A* are *not* identical: let us say that W_1 is one of those worlds in which an atomic war was fought in 1966. Surely, if there is such a possible world,[4] it would be odd to say that anything that could reasonably be called "determinism" is true. In the case of *A* we have a world in which a certain situation in 1066 is *not* followed, nine hundred years later, by an atomic war.

[4] Of course, "a possible world distinct from the actual world, and bearing both *N* and *S* to it" is not a self-contradictory description, but it does not follow that there is any possible world answering to it. Similarly, "a possible world in which the first assertion made by Richard Nixon in the actual world in 1972 holds true" is not a self-contradictory description, but there may be no possible world answering to it.

But in W_1, a world having exactly the same laws of nature, precisely the same situation is followed, after nine hundred years, by an atomic war. In other words, though there was no atomic war in 1966, such a war was a *possibility* relative to the laws of nature and the state of the world in 1066. But surely "determinism" must, if violence is not to be done to every traditional association that word has, be used to refer to some thesis according to which there are no such alternate possibilities. Let us, therefore, understand by "determinism" the thesis that the actual world is deterministic.[5]

One might want to ask at this point whether determinism in this sense is true or false according to the usual interpretations of quantum mechanics. The answer seems to me to be that it is false. According to these interpretations, there can be two unstable atomic nuclei (neither of which is subject to any external influence) that are in *exactly* the same state at some instant, and which decay at different times. If that is the case, it is easy to imagine a possible world nomologically congruent to the actual world and indistinguishable from it at one instant, but distinguishable from it at some later instant.

One might also want to ask whether a purely Newtonian possible world (a world of point-masses behaving in accordance with Newton's laws of motion and the law of universal gravitation) would be deterministic. This is a difficult question to answer. For a two-particle Newtonian world the answer is certainly Yes: there is only one possible two-particle Newtonian world relative to any specification of boundary conditions, since the differential equations describing such a world have a *unique* general solution.

[5] This notion of determinism derives from the model-theoretic concepts of a "deterministic theory" and a "deterministic history" developed by Richard Montague in "Deterministic theories," in *Decisions, values and groups*, ed. by N. F. Washburne (New York: Pergamon Press, 1962). I am indebted to Rolf Eberle for calling my attention to this important paper, and for allowing me to attend a seminar at the University of Rochester in which he gave a lucid exposition of it. A notion of determinism very similar to the one presented in this paper, and also based on Montague's work, is presented by John Earman in his 1971 A.P.A. Symposium paper, "Laplacian determinism, or Is this any way to run a universe?" printed in *The journal of philosophy*, vol. 68 (1971), pp. 729—744.

The question whether in general a Newtonian *n*-particle world, where $n > 2$, is deterministic is at present unanswered, since it is not known whether there is a general and unique solution to the appropriate differential equations.[6]

The predicate "*H*" represents the relation that *x* bears to *y* if and only if *x* is an actual person (i.e., an *actuale* of the sort that deliberates about future courses of action) and *y* is a possible world and *x* has access to *y*. In order to clarify what is meant by saying that a person "has access to" some world distinct from the actual world (we may take it to be true by definition that everyone has access to the actual world), I shall first give some translations from ordinary talk about *abilities* to "access" talk. I do not say that the translations have the same meanings as the originals. I am claiming only that the translations could be used in place of the originals, and for the same purposes. We translate, "Napoleon could have defeated Wellington at Waterloo" as, "Napoleon had access to some possible world in which Napoleon defeated Wellington at Waterloo." We translate, "It is within my power to keep the money I found and within my power to return it" as, "I have access to at least one possible world in which I keep the money I found and to at least one possible world in which I return it."

The following bit of dialogue indicates how our moral discourse might sound if we gave up ordinary ability-talk, and adopted in its place the language of access to possible worlds:

A. You ought not to have cut my lecture on Friday.

B. But I had no access to a possible world in which I attended your lecture on Friday, since I suffered an unforeseen paralysis of my legs on Thursday that mysteriously vanished on Saturday. In every possible world to which I had access, I spent Friday in bed.

A. Have you access to a possible world in which a doctor writes me a note verifying your story?

B. Unfortunately not: no possible world to which I had access Friday contained a doctor in this city who makes house calls.

[6] Cf. Montague, op.cit., p. 349 ff.

And so on. Perhaps the relationship between ordinary ability-talk and access-talk might best be explicated by showing the relationship between access-talk and a rather artificial near-relation of ordinary ability-talk, viz., talk of one's abilities with respect to bringing about events of some specified sort: to say that a person can bring about an event satisfying a certain description is to say that he has access to at least one possible world in which an event satisfying that description happens; and to say that a person has access to a possible world satisfying a certain description is to say that he can bring about events of a sort that happen only in worlds satisfying that description.

In order to make this relationship intuitively more clear, I shall devise a sort of metaphor or "picture" that might be used as an informal model both for talk of being able to bring about events and talk of access to possible worlds. Consider a man who is walking through an infinite system of branching corridors. He has always been walking and must always keep walking, never stopping and never retracing his steps. He finds that some branches are sealed off by bars and some are not. Frequently he comes to a branching of the corridor from which at least two unbarred branches lead away, and he must make a choice about which to take.

Let us call any location within the system of corridors an *event*. Then we may say that the man *can bring about* a certain event just in the case that there is some path through the corridors from where he is to that event (location) that does not lead through any barred corridors.[7]

Let us call a *possible world* any infinitely long path through the system of corridors that does not cross itself. The *actual world* is that one path through the corridors along which the man always has walked, is walking, and always will walk. Those worlds to which the man *has access* at any given moment are just those

[7] The bars are, of course, as much a piece of imagery as the system of corridors. My use of them in this model is not meant to suggest that an agent is unable to bring about an event only in the case that some tangible and immovable barrier stands between him and the means necessary for bringing it about.

infinite paths that are continuations of the path-segment along which he has already walked that do not pass through any barred corridors.

This "picture" has its limitations as a model for talk of access to possible worlds: it is no longer applicable if we assume (as is the case) that *which* possible world the actual world is depends on the choices of more than one person. We might, of course, elaborate our imagery by assuming that there are *n* persons walking through the system of corridors, and call a *possible world* any *n*-membered *set* of infinite paths. The actual world, then, would be the set of paths that *are* taken, and a person *P* would have access at any given moment to those possible worlds that are such that (i) they differ from the actual world by at most one member, (ii) this member is the path that *P* is in fact going to take, and (iii) each of them that does not contain the path that *P* is in fact going to take, contains instead a continuation of the path-segment *P* has already walked that does not pass through any barred corridors.

But this more elaborate picture breaks down in its turn if we assume (as is the case) that persons come into and go out of existence, and that the choices they make partly determine what choices it is *possible* for their fellows to make. I do not think, however, that there is anything to be gained from constructing a yet more elaborate picture in order to accomodate these facts.

We should note that *H* is, strictly speaking, a non-temporal relation between persons and possible worlds: it is not a triadic relation satisfied by ordered triples of the form ⟨person, world, instant⟩, but a dyadic relation satisfied by ordered pairs of the form ⟨person, world⟩. For example, if Tom, a doctor, once had access to a possible world W_2 in which his profession is law, then, even if he no longer has access to W_2, it is true that Tom bears *H* to W_2. Thus, a better English reading of "*Hxy*" might be "*x* had, has, or will have access, at some point in his life, to *y*."

II

The thesis I shall call the *minimal free-will thesis* (MFT) may be expressed formally as:

$$\exists x \exists y(Hxy \;\&\; y \neq A).$$

That is to say, *some* person (past, present, or future) had, has, or will have access to *some* possible world besides the actual world. This is a very weak thesis. It is true, for example, if Julius Caesar bore H to some possible world W_3 in which he did not cross the Rubicon, even if no other person, past, present, or future, bears H to anything besides A, and Caesar himself bore H only to W_3 and A. But if the minimal free-will thesis were false, then, surely, any more interesting free-will thesis would be false.

Let us now ask whether determinism logically entails the denial of the minimal free-will thesis, or, more precisely, whether the negation of MFT is deducible from "DA". It is clear by simple inspection that the answer to this question is No. Nevertheless, there is an important sense in which the truth of determinism insures the falsity of the minimal free-will thesis: there are two theses, which I shall call "metaphysical assumptions," each of which seems more likely to be true than either determinism or the minimal free-will thesis, and such that the denial of the minimal free-will thesis follows logically from determinism and these two theses taken together. The two metaphysical assumptions are:

MAA $\forall x \forall y(Hxy \supset NyA).$
MAB $\forall x \forall y(Hxy \supset SyA).$

If we read "Nxy" as, "the laws of nature are the same in x and y," then MAA asserts that no person has access to any world in which the laws of nature are different from what they are in the actual world. This seems undeniable. What the laws of nature *are* does not depend upon human choice, though, of course, our *beliefs* about what statements are most probably laws of nature may. For example, it may be that if some physicist had performed a certain experiment (which he would have performed if he had not thought some other line of inquiry more promising), then we should now believe the principle of the conservation of linear momentum to be false. But if this were true we should not say that the physicist had access to a possible world in which the laws of nature were different from the actual laws, but (at most) that he had access to a possible world in which our *conception* of the laws of nature was different from our actual conception.

MAB asserts that every world to which any person has access must be indistinguishable from the actual world at some point in time. Or, alternatively, every world to which any person has access must share a slice with the actual world. For example, however many possible worlds I have access to, surely they must all be indistinguishable from the actual world at some time in the remote past (say, 10,000 B.C., or, indeed, any time before I was born). In terms of the "infinite-system-of-corridors" metaphor: all the possible worlds (paths) that I have access to are continuations of the path-segment I have already traveled. MAB is a rough echo of the familiar principle that no one can change the past.

I shall now present an informal proof of the negation of MFT. The only assumptions made will be "*DA*", MAA, and MAB. The proof is trivial and could easily be made rigorous. Assume:

$$(1) \qquad Hxy.$$

From (1) and the universal instantiation of MAA:

$$(2) \qquad NyA.$$

Similarly, from (1) and MAB:

$$(3) \qquad SyA.$$

From (2), (3) and "*DA*":

$$(4) \qquad y = A.$$

And by conditional proof and universal generalization:

$$(5) \qquad \forall x \forall y (Hxy \supset y = A),$$

which is logically equivalent to the denial of MFT.

In this sense, then, determinism and free will are incompatible: assuming "*DA*", MAA, and MAB we may deduce the negation of MFT. And MAA and MAB are undeniable truths. They can, of course, be rejected without formal contradiction, but I do not find their denials very intelligible. What could it mean to say that someone has access to a possible world in which the laws of nature are different from our laws, or to a possible world having a different history from ours? In particular, how could we understand a man who claimed to have access to a possible world in which the speed of light is twenty miles per hour, or to a possible world in which Lincoln lived to be eighty years old? I submit that if we suppose that he *understands* the claims he is making, then we can only suppose that he is grossly mistaken about the facts: he must believe that the speed of light *is* twenty miles per hour, or that it varies in accordance with some natural law that he can exploit; or he must believe that Lincoln *did* live to be eighty years old, or that Lincoln is alive and less than eighty years old. If he agrees with us that the speed of light is much greater than twenty miles per hour and is fixed as a matter of natural law, and if he agrees with us that Lincoln died over a hundred years ago at an age considerably less than eighty, then we can only suppose that he does not understand the claim he is making.

It is important to realize that the soundness of our argument does not depend on what the correct answer is to the question whether abilities are "hypothetical" or "categorical." For we might interpret the notion of access to a possible world hypothetically: we might define "x has access to a possible world satisfying description Φ" as meaning something like, "if x were to choose to bring it about that the actual world satisfies Φ, then the actual world would satisfy Φ." But this definition could be used to show that our argument is unsound only if it could be used to show that at least one of our two metaphysical assumptions is false. And this does not seem to be the case: if Φ_1 is a description that applies only to worlds that are not nomologically congruent to the actual world, e.g., "containing moving material objects that make perfectly sharp right-angle turns," then no choice of mine could bring it about that the actual world satisfies Φ_1; and if Φ_2 is a description that applies only to worlds in which the past is different from the actual past, e.g., "in which wireless telegraphy was invented in 1850," then no choice of mine could bring it about that the actual world satisfies Φ_2.

III

Given that determinism and the minimal free-will thesis are inconsistent, which ought we to reject? One simple reply is that we should reject determinism since it is incompatible with currently accepted physical theory. But if a mistake should be found in von Neumann's argument against the possibility of introducing "hidden parameters" into standard quantum theory in such a way as to make it deterministic, or if physics should undergo some unforeseen radical transformation, then we might be faced with the problem again, and it seems best to ask what we should say in these cases. Moreover, since it seems unlikely that the macroscopic movements of human bodies normally depend on individual events on the quantum level, it might be possible to devise some empirically tenable theory, B-determinism, according to which a human body is a kind of deterministic subsystem of a world that is, taken as a whole, indeterministic. And if determinism can be shown to be incompatible with the minimal free-will thesis, it seems reasonable to suppose that a similar proof could be devised for the incompatibility of MFT and B-determinism. Therefore, it should seem, the simple answer suggested above is little more than an evasion of the real issue.

The question whether we should reject determinism or reject the minimal free-will thesis (once we have decided that they are incompatible) is a profound and difficult question to which I do not know the answer. I would, however, suggest that anyone who attempts to answer it consider carefully the following two points. (1) There seems to be no reason to think that determinism is a "presupposition of science." We see this not only in the case of a "statistical" physical theory like quantum mechanics, but even in the case of classical celestial mechanics, the paradigm of a successful predictive science. If it could be shown that there is no unique general solution to certain systems of simultaneous differential equations, this would suffice to show that (typical) Newtonian worlds containing more than two particles are not deterministic in our sense. But such a mathematical discovery would make no difference to the practice of the science of celestial mechanics. (2) One reason philosophers have been reluctant to discard principles similar to the minimal free-will thesis is that these principles are commonly thought to be presupposed by ascriptions of moral responsibility. But the principle that if the minimal free-will thesis is false, then no one is morally responsible for his acts is very much like what Harry Frankfurt has called the "principle of alternate possibilities," a principle that Frankfurt claims is false.[8] While I think that Frankfurt fails to show conclusively that the principle of alternate possibilities is false, I think that this principle (or family of related principles) does not deserve the uncritical endorsement it has had from most moral philosophers. Certainly any argument that, though we are unable to decide on *factual* grounds whether determinism is true or false, we are nonetheless justified in rejecting it on *practical* grounds (i.e., in order to allow for moral responsibility), is premature, even if we grant that determinism and free will are incompatible: this argument presupposes the principle of alternate possibilities, which is in urgent need of clarification and analysis.[9]

[8] "Alternate possibilities and moral responsibility," *The journal of philosophy*, vol. 66 (1969), pp. 829—839.

[9] I wish to thank Mark Brown and Rolf Eberle for helpful comments on an earlier draft of this paper, and Keith Lehrer for valuable discussions of the problem of free will and determinism which have had considerable, if indirect, influence on this paper. This paper was read at the Summer Institute in Philosophical Theology (Calvin College, Grand Rapids, 1973). I wish to thank the members and staff of the Institute, and, in particular, George Mavrodes and Alvin Plantinga, for acute and useful criticism.

Received on January 11, 1973.