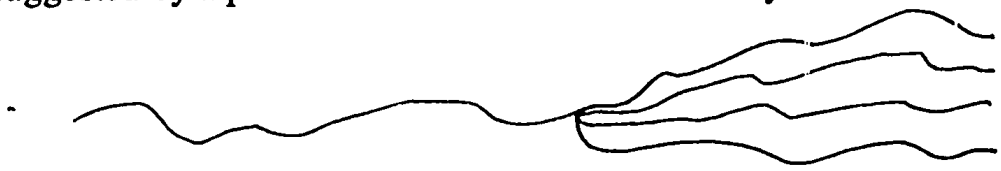
Logic and the Free Will Problem

The best way to get an intuitive grip on the problem of free will and determinism is to think of time as a "garden of forking paths." That is, to think of the alternatives that one considers in deliberation as being incidents that are included in "alternative futures," and to think of alternative futures diagrammatically, in the way suggested by a path or a river or a road that literally forks:



If Jane is trying to decide whether to tell all or to continue her life of deception, she is in a situation strongly analogous to that of someone who is hesitating between forks in a road. That is why this sort of diagram is so suggestive. Let us see how the idea of time as a garden of forking paths helps us to understand the problem of free will and determinism.

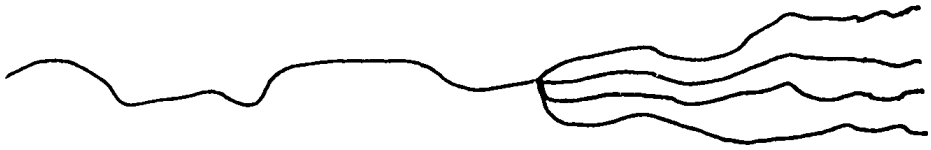
To say that one has free will is to say that when one decides among forks in the road of time (or, more prosaically, when one decides what to do),[1] one is at least sometimes able to take more than one of the forks, Thus, Jane, who is deciding between a fork that leads to telling all and a fork that leads to a life of continued deception, has free will (on this particular occasion) if she is able to tell all and is also able to continue living a life of deception. One has free will if sometimes more than one of the forks in the road of time are "open" to one. One lacks free will if on every occasion only one of the forks—of course it will be the fork one in fact takes—is open to one. If John is locked in a room and doesn't know that he is locked in, and if he is in the process of deliberating

about whether to leave, one of the alternative futures he is contemplating—leaving—is, in point of fact, not open to him, and he thus lacks free will in the matter of staying or leaving.

It is a common opinion that free will is required by morality. Let us examine this common opinion from the perspective that is provided by looking at time as a garden of forking paths. While it is obviously false—for about six independent reasons—that the whole of morality consists in making judgments of the form, "You should not have done x," we can at least illustrate certain important features of the relation between free will and morality by examining the relation between the concept of free will and the content of such judgments. To say that you shouldn't have done x is to say that you should have done something else instead. To say that you should have done something else instead is to say that you *could* have done something else. (If this were not true, one could say, "I concede that you were unable to do x instead of what you did. I nevertheless maintain that you should not have done what you did, but should have done x instead.") To say that you could have done something else is to say that you have free will. To make a moral judgment about a person's act is to evaluate his taking one of the forks in the road of time relatively to one or more of the forks that were also open to him. (Note that if John makes a choice by taking one of the forks in what is literally a road, one cannot blame him for taking the fork he did if all of the other forks were blocked.) A moral evaluation of what someone has done requires two or more alternative possibilities of action for that person, just as surely as a contest requires two or more contestants.

Let us now see what help the conception of time as a garden of forking paths gives us in understanding determinism. Determinism is the thesis that it is true at every moment that the way things then are determines a unique future, that only one of the alternative futures that may exist relative to a given moment is a physically possible continuation of a state of things at that moment. Or, if you like, we may say that determinism is the thesis that only one continuation of the state of things at a given moment is consistent with the laws of nature. Thus, according to determinism, although it may often seem to us that we confront a

sheaf of possible futures (like this)



what we really confront is something like this

Here the dotted lines represent futures that are not physically possible continuations of the present, and the single solid line represents the future that the laws of nature permit.[2]

It has seemed obvious to most people who have not been exposed (I sometimes think that "subjected" would be a better word) to philosophy that free will and determinism are incompatible. It is almost impossible to get beginning students of philosophy to take seriously the idea that free will and determinism are compatible. Indeed, people who have not been exposed to philosophy usually understand the word "determinism" (if they know the word at all) to stand for the thesis that there is no free will. One might think that the incompatibility of free will and determinism deserves to seem obvious—because it is obvious. To say that we have free will is to say that more than one future is sometimes open to us. To affirm determinism is to say that every future that confronts us but one is physically impossible. And surely a physically impossible future can't be open to one, can it? If we know that a "Star Trek" sort of future is physically impossible (because, say, the "warp drive" that figures essentially in such futures is physically impossible), then we know that a "Star Trek" future is not open to us or to our descendants.

People who are convinced by this sort of reasoning are called *incompatabilists*. As I have hinted, however, many philosophers are *compatibilists*: they hold that free will and determinism are compatible. Compatibilism has an illustrious history, which embraces such figures as Hobbes, Hume, and Mill, and no doubt the majority of present-day philosophers, at least in the English-speaking countries, are compatibilists.[3] A modern

compatibilist can be expected to reply to the line of reasoning I have just presented in some such way as follows. "Yes, a future, in order to be open to one, does need to be physically possible. It can't, for example, contain faster-than-light travel if faster-than-light travel is physically impossible. But we must distinguish between a future's being physically possible and its having a physically possible connection with the present. A future is physically possible if it is permitted by the laws of nature. A future has a physically possible connection with the present if it could be 'joined' to the present without any violation of the laws of nature. A physically possible future that does not have a physically possible connection with the present is one that, given the present state of things, would have to be 'inaugurated' by an event that violated the laws of nature, but in which, thereafter, events proceed in accordance with the laws. Determinism indeed says that of all the physically possible futures, one and only one has a physically possible connection with the present—one and only one could be joined to the present without a violation of the laws of nature. My position is that some futures that could not be joined to the present without a violation of the laws of nature are, nevertheless, open to us."

Two philosophical problems face the defenders of compatibilism. The easier is to provide a clear statement of which futures that do not have a physically possible connection with the present are "open" to an agent. The more difficult is to make it seem at least plausible that futures that are in this sense open to an agent really deserve to be so described.

An example of a solution to these problems may make the nature of the problem clearer. The solution I shall briefly describe would almost certainly be regarded by all present-day compatibilists as defective, although it has a respectable history. I choose it not to suggest that compatibilists can't do better, but simply because it can be described in fairly simple terms.

According to this solution, a future is open to an agent if, given that the agent chose that future (chose that fork in the road of time), it would come to pass. Thus it is open to me to stop reading this paper and do a little dance, because if I so chose, that is what I would do. On the other hand, if Alice is locked in a prison cell, it

is not open to her to leave: if she chose to leave, her choice would be ineffective because she would come up against a locked prison door. Now consider the future I said was open to me—to stop reading and do a little dance—and suppose that determinism is true. Although a choice on my part to entertain you in that remarkable fashion would (no doubt) be effective if it occurred, it is determined that such a choice will not occur. It is in fact determined that nothing is going to occur that would have the consequence that I stop reading and do a little dance. Therefore, none of the futures in which I behave in this bizarre fashion is a future that has a physically possible connection with the present: every such future could come to pass only if it were inaugurated by an event of a sort that is ruled out by the present state of things and the laws of nature. And yet, as we have seen, many of these futures are "open" to me in the sense the compatibilist has proposed. Is this a reasonable sense to give to this word? (We now take up the second problem that confronts the compatibilist.) This is a very large question. The core of the compatibilist's answer is an attempt to show that the reason we are interested in open or accessible futures is that we are interested in modifying the way people behave. One important way in which we modify behavior is by rewarding behavior that we like and punishing behavior that we dislike. We tell people that we will put them in jail if they steal and that they will get a tax break if they invest their money in such-and-such a way. But there is no point in trying to get people to act in a certain way if that way is not in some sense open to them. There is no point in telling someone that he will go to jail if he steals unless it is somehow open to him not to steal. And what is the relevant sense of "open"? Just the one I have proposed, says the compatibilist. One modifies behavior by modifying the choices people make. That procedure is effective just insofar as choices are effective in producing behavior. If someone chooses not to steal (and remains constant in that choice), then he won't steal. On the other hand, if someone chooses not to be subject to the force of gravity, he will nevertheless be subject to the force of gravity. Although it would no doubt be socially useful if there were some people who were not subject to the force of gravity, there is no point in threatening people with grave consequences if they do

not break the bonds of gravitation, for even if one managed to induce some people to choose not to be subject to the force of gravity, their choice would not be effective. Therefore (the compatibilist concludes), it is entirely appropriate to speak of a future as "open" if it is a future that would be brought about by a choice. And if someone protests when you punish him for not choosing a future that was in this sense open to him, on the ground that it was determined by events that occurred before his birth that he not make the choice that would have inaugurated that future, you can tell him that his punishment will not be less effective in modifying his behavior (and the behavior of those who witness this punishment) on *that* account.

What I have tried to do so far is to give a brief and intuitive account of free will and determinism, and of the two major approaches to the problem, the approaches taken by the incompatibilist and the compatibilist. What I want to do in the sequel is to look at one aspect of the problem in some detail. I am afraid it must be rather technical detail, for there comes a point in the discussion of a philosophical problem at which one can go no further without going into technical detail. I am going to present an argument for the incompatibility of free will and determinism, and defend this argument against an objection that many philosophers have found cogent.[4]

Let 'N' be an operator that expresses whatever sort of necessity it is that is opposed to free will. It seems plausible to suppose that the following two inference-rules governing 'N' are valid:

(α) $\Box p \ \vdash Np$ (where '$\Box$' has its standard sense)

(β) $Np, N(p \supset q) \ \vdash Nq$.

Now let 'S' be a sentence that gives a complete description of the state of the world at some time in the remote past. Let 'L' be a sentence that expresses the conjunction into a single proposition of the laws of nature. Let 'T' be any sentence that expresses a truth about the present or the future. Now assume determinism. It is a consequence of determinism that

$\Box((S \cdot L) \supset T)$.

And from this consequence of determinism, it follows, by elementary modal and sentential logic, that

$\Box(S \supset (L \supset T))$.

From this it follows by (α) that

$N(S \supset (L \supset T))$.

Now it would seem that both the laws of nature and statements about the past—and certainly the remote past—should be accounted "necessary" in the sense expressed by 'N'.[5] We have, therefore 'NS' and 'NL', and then, by two applications of (β), 'NT'. That is, if determinism is true, then any truth whatever is "necessary" in a sense that is opposed to our having free will about whether it is true. If the argument that has led to this conditional is valid, then determinism is incompatible with free will.

Is the argument valid? Since no one (I think) would want to dispute (α), this question reduces to the question whether (β) is valid. And that is a very good question indeed.

In an important and much-cited article, Michael Slote has attempted to cast doubt on the validity of (β).[6] Slote suggests that anyone who accepts (β) probably accepts it only because he accepts these two rules:

Agglomeration $Np, Nq \mid\!- N(p \cdot q)$

Closure $Np \mid\!- Nq$, provided q is derivable from p.

He says:

> Anyone who assumes the validity of arguing from 'N*p*' and 'N ($p \supset q$)' to 'N*q*' would seem to be tacitly assuming that the necessity expressed in the operator 'N' is both agglomerative (closed with respect to conjunction introduction) and closed under logical implication, so that one can, for example, validly move from 'N*p*' and 'N ($p \supset q$)' to 'N ($p \cdot p \supset q$)' and from the latter to 'N*q*'. *If we do not think about these subinferences, when we move from 'Np'* and 'N($p \supset q$)' to 'N*q*' or assert the...modal principle [(β)] that corresponds to that larger inference, that is only because it is so natural to assume that any necessity

operator will have the properties of agglomerativity and closure under logical implication or entailment.

If I understand my own allegiance to (β), however, it does not seem to have come about in this way. Let me try to explain this allegiance. To begin with, I believe that the necessity that (as I rather vaguely put it above) "is opposed to free will" should be spelled out in this way:

Np = df p and no one has, or ever had, any choice about whether p.

At any rate, an argument for the conclusion that determinism entails that every true proposition is necessary in this sense is certainly correctly describable as an argument for the incompatibility of free will and determinism. (Moreover, the premises of the argument for the incompatibility of free will and determinism that I laid out above—'NS' and 'NL'—would seem to be obviously true on this interpretation of 'N', at least if we interpret 'S' as a statement about the state of the world before there were any beings of the sort whose free will we are interested in.)

I believe that (β) is valid if the operator 'N' is interpreted in this way. (Let us say that on this interpretation, 'N' expresses *Choice Necessity*.) But, so far as I can tell, this conviction of mine does not arise from a prior conviction, or even a tacit assumption, that Agglomeration and Closure, or any other inference-rules, are valid.[7] I believe that this conviction arises from the intrinsic plausibility of (β) when 'N' is interpreted as expressing Choice Necessity. (In the sequel, 'N' is to be understood as expressing Choice Necessity—and (β) is to be interpreted accordingly—unless I explicitly stipulate some other interpretation.) I cannot, in the strictest sense of "argument," give an argument for the thesis that (β) is valid, for I know of no thesis more plausible than itself from which it follows. But I may be able to communicate my conviction that (β) is valid, and communicate it in a very simple and straightforward way. If you wish to appreciate the plausibility of the validity of (β), attempt to construct a counterexample to it. I believe that the reader who has

made a serious and sustained effort to construct a counterexample to (β) will come to share my conviction that (β) is valid, or, if not to share it, then at least to see how someone could have this conviction quite independently of his convictions about the validity of any other rules of inference.[8]

I will remark in passing that I *am* inclined to think that Agglomeration and Closure are valid. Consider the set of worlds W such that the actual world belongs to W and a non-actual world belongs to W if and only if someone has, or once had, a choice about whether that world is actual. It seems to me plausible to suppose that W is such that, for any *p*,

p and no one has, or ever had, any choice about whether *p*

is true (*sc.* in the actual world) if and only if *p* is true in every member of W.[9] That is to say, there is a certain set of worlds (containing the actual world) such that for every *p*, ⌜N*p*⌝ is true if and only if *p* is true in every member of that set. More generally, for every world *w*, there is a set of worlds W (containing *w*) such that for every *p*, ⌜N*p*⌝ is true in *w* if and only if *p* is true in every member of W. This means that 'N' is what we might call a "classical" necessity operator and not what Slote calls a "selective" necessity operator. I leave as an exercise the trivial proof that Agglomeration and Closure hold for all classical necessity operators.

I am therefore inclined to think that 'N' is, as Slote puts it, "both agglomerative and closed under logical implication" because I am inclined to think that the following two operators are equivalent:

p and no one has, or ever had, any choice about whether *p*

The proposition that *p* is true in the actual world and in all non-actual worlds such that someone has, or once had, a choice about whether they are actual.

But if someone could convince me that this equivalence did not hold, and could, in fact, convince me of the truth of the stronger statement that the Choice Necessity operator was not a classical necessity operator, I should still accept (β). (We should note that

while the statement that 'N' is a classical necessity operator entails the validity of (β), the validity of (β) does not entail that 'N' is a classical necessity operator.[10]) I should still accept (β) because my conviction that (β) is valid rests on what I believe I have learned by attempting to construct counterexamples to (β) and not on my belief that the two sentences displayed above are equivalent or on my belief that the Choice Necessity operator is a classical necessity operator. (My belief that there are people does not rest on my belief—or on my tacit assumption—that there are people in Tibet, despite the fact that I do believe, and with great conviction, that there are people in Tibet, and know that the proposition that there are people in Tibet entails the proposition that there are people. If something convinced me that there were, after all, no people in Tibet, I should still believe that there were people.)

I have implied that it is at least very difficult to find a counterexample to (β). I will now say this explicitly, and while I am at it, make explicit an important qualification: it is at least very difficult to find a counterexample to (β) that can be seen to be a counterexample independently of the question whether free will is compatible with determinism. Of course if free will is compatible with determinism, it is easy to find a counterexample to (β); in fact we have already done so. If free will is compatible with determinism, then one or the other of the two applications of (β) in the argument for incompatibilism displayed above must have taken us from truth to falsity. But, of course, putative counterexamples to (β) that "work" only if compatibilism is true are of no interest in a dispute about the truth of compatibilism. What would be of interest would be a putative counterexample to (β) that could be evaluated independently of the question whether free will and determinism were compatible.[11]

My allegiance to (β), therefore, is quite independent of whatever opinions I may hold about Agglomeration and Closure, and an argument that led me to doubt one or both of these principles could very well leave my allegiance to (β) unshaken. (I cannot deny that an argument that led me to doubt Agglomeration or Closure could also be an argument that undermined my allegiance to (β). After all, any argument that did undermine my allegiance to (β) would almost certainly also be an argument that led me to doubt

Agglomeration or Closure—since the invalidity of (β) entails that either Agglomeration or Closure is invalid.) Does Slote say anything that one might apply "directly" to (β)? Only this. He displays various "plausible instances of alethic necessity" for which (β) fails—for example, non-accidentality, irresistible impulse, and compulsion.[12] These certainly seem to be in some sense types of "necessity," and Slote shows convincingly that (β) is not valid if "N" is interpreted as expressing any of them. But what is the point of this procedure? I concede that it might serve to undermine an allegiance to (β) that was based on a belief that all necessity operators were "classical," or was based on an unexamined analogy between Choice Necessity and standard logical or metaphysical (or even physical) necessity. But I do not know of anyone whose allegiance to (β) *does* rest on so infirm a foundation. I know that mine doesn't, and I suspect that Ginet and Lamb and Wiggins would say the same.

I am moreover, puzzled that Slote should bother discussing these relatively uninteresting selective necessity operators when it is easy to find selective necessity operators that seem to be far more relevant to the problem of free will and determinism. Consider this one, for example:

> p and no one is, or ever has been, such that if he were to choose to bring it about that it is (or was) false that *p*, then it would be (or would have been) false that *p*.

The rule (β) fails for this operator.[13] Moreover, this operator is of special interest in discussions of the free will problem because adherents of the popular view that "*x* can do A" is equivalent to "If *x* were to choose to do A, *x* would do A" would, presumably, say that it was equivalent to our Choice Necessity operator:

> *p* and no one has, or ever had, any choice about whether *p*.

If this thesis—call it the Equivalence Thesis—is correct, it follows that our argument for the incompatibility of free will and determinism is invalid. I am not greatly troubled by this; in fact, matters could hardly be otherwise. It is obvious that if "can" statements

are, as so many compatibilitists allege, a certain sort of disguised conditional, then any argument for the incompatibility of free will and determinism is either invalid or else has false premises. It is obvious that if the Equivalence Thesis is correct, then "can" statements are disguised conditionals of that type. It is obvious that if our argument is invalid it is only because (β) is invalid. It is therefore not surprising that, if the Equivalence Thesis is correct, then (β) is invalid. So much the worse for the Equivalence Thesis, I say. It is obvious that (β) is valid (or so it seems to me); it is not obvious that the Equivalence Thesis holds. If two propositions are incompatible and one seems obviously true and the other does not seem obviously true, then, all other things being equal, one should accept the obvious member of the pair if one accepts either.

I am, therefore, unmoved by the fact the (β) fails if 'N' is interpreted as expressing the "were to choose" operator. But why then should I be moved by the fact that (β) fails if 'N' is interpreted as expressing the operators ("it is no accident that *p*" and so on) that Slote calls our attention to? I do not think that the "were to choose" operator expresses Choice Necessity. But it is certainly true that many philosophers have thought that it does, or have held views that entail that it does. Therefore, there is an intimate and important connection, at least in the minds of many philosophers, between the "were to choose" operator and the problem of free will and determinism. The operators that Slote displays, however, are much less intimately connected with the problem of free will and determinism, and the fact that (β) fails for these operators is consequently even less troubling to the incompatibilist than the fact that (β) fails for the "were to choose" operator.

Notes

1. The term "free will," when used in this sense, is a term of art. To ascribe "free will" to an agent is simply to ascribe to the agent the property of having a free choice among certain alternatives. Such an ascription should not be taken to imply that the agent has a faculty called "the will."
2. During the discussion of this paper, David Lyons suggested that the following diagram might be more perspicuous.

This diagram has the following interesting feature: if it is viewed from a distance, the viewer will not be able to tell which of the forks is the one that is continuous with the past. This may be thought of as representing the fact that, from our point of view (which is not that of God or the Laplacian Intelligence), we cannot tell which member of a sheaf of possible futures is the one that is continuous with reality.

3. Professor Slote, in his comments, contends that the situation has changed sufficiently in the last few years that it is no longer clear that the majority of English-speaking philosophers are compatibilists. On reflection, I am inclined to agree with him.
4. This argument is discussed in greater detail than is here possible (or necessary) in Chapter III of my book *An Essay on Free Will* (Oxford: The Clarendon Press, 1983).
5. "Not even the gods can change the past"; *only* a God—or, at least, some sort of supernatural being—could change (or violate, set aside, defy, cause things to act in ways other than those prescribed by) the laws of nature.
6. "Selective Necessity and the Free-Will Problem," *The Journal of Philosophy*, 79 (1982); 5-24. Slote's article is a criticism of four defenses of the incompatibility of free will and determinism: Carl Ginet, "Might We Have No Choice?" in Keith Lehrer, ed., *Freedom and Determinism* (New York: Random House, 1966), pp. 87-104; James Lamb, "On a Proof of Incompatibilism," *Philosophical Review* 86 (1977); 20-35; David Wiggins, "Towards a Reasonable Libertarianism," in Ted Honderich, ed., *Essays on Freedom of Action* (London: Routledge and Kegan Paul, 1973), pp. 31-61; my "The Incompatibility of Free Will and Determinism," *Philosophical Studies* 27 (1975): 185-99. My article is reprinted in Gary Watson, ed., *Free Will* (New York: Oxford University Press, 1982), pp. 46-58.
   Slote has a certain amount of trouble with my article, the argument of which is laid out in a way that is rather inconvenient for his purposes. But he is perfectly right in thinking that the points he makes are as applicable to my article as the other three. Because the argument of Section 3.10 of my book (see *n.* 4) is laid out in such a way that Slote's points can be applied to it without trivial and annoying adjustments and qualifications, I shall defend that argument rather than the argument of "The Incompatibility of Free Will and Determinism."
7. Nor, as Slote repeatedly suggests, does the example of standard alethic modal logic play any role in my conviction. Or not so far as I can tell.
8. I am not offering an inductive argument. I am not proposing that the reader, having made (say) sixteen failed attempts to construct a counterexample to (β), should reason as follows: "All sixteen of my attempts to find a counterexample to (β) have been failures. Therefore, probably, (β) is valid." Rather, my hope is that the reader, in the course of attempting to construct counterexamples to (β), will come to *see* why (β) is valid.

Here is a general description of what one experiences when one attempts to construct a counterexample to (β). One substitutes particular declarative sentences for '*p*' and '*q*', and one devises a case according to which, one hopes, the premises of the resulting argument are true and its conclusion false. On careful examination, however, it transpires that (say) one of the premises is false; one adjusts the case to make the premise true and thereupon discovers that one has inadvertently made the other premise false; a second adjustment, intended to correct this fault, causes the conclusion to be true; when *this* defect is corrected, it turns out that the premise that occasioned the original modification of the case is once more false. As the hours pass—hours during which one tries many and various pairs of substitutions for '*p*' and '*q*' and constructs many bizarre scenarios—one begins to recognize patterns in the repeated blockings of the "TTF" case, patterns that display to one the inevitability of the frustration of every attempt to devise an instance of that case.

9. I shall not invariably be careful about distinguishing use and mention or variables from dummy letters.
10. For example, let W be any possible world. If 'N' is interpreted as 'it is true in W that', (β) is valid and N is not a classical necessity operator (as is easily seen from the fact that *p* could be false and N*p* true).
11. The best I know of is due to Thomas McKay. Suppose John has a choice about whether he plays dice, and, in fact does play dice. But no one has a choice about how the dice fall in a fair game. The game *is* fair. John throws a six. The proposed counterexample is:

    N John throws a six
    N (John throws a six ⊃ John plays dice)
    ∴N John plays dice.

    The first premise is true for the reason given. The second premise is true because the embedded conditional is a necessary truth. The conclusion is false for the reason given.

    I reply: the first premise *is* false. John could have avoided throwing a six by avoiding playing dice. What John has no choice about is whether he throws a six given that he plays dice; that is, about whether (John plays dice ⊃ John throws a six).
12. I'm not sure what Slote means by "alethic." In a draft of this paper, I suggested "neither deontic nor doxastic," but Slote has denied (in correspondence) that this was what he meant. He did not, however, explain what he did mean.
13. See *An Essay on Free Will*, p. 122ff.

Peter van Inwagen
Department of Philosophy
Syracuse University