Philosophical Perspectives, 3
Philosophy of Mind and Action Theory, 1989

WHEN IS THE WILL FREE?

Peter van Inwagen
Syracuse University

There is, it seems to me, something that might be called an "orthodox" or "classical" tradition in the history of thinking about the problem of free will and determinism. This tradition, as I see it, descends from Hobbes through Locke and Hume and Mill to the present day. I say "it seems to me" and "as I see it" because I am no historian and I freely grant that what appears to my untutored mind to be "the classical tradition" in the debate about free will and determinism may be an artifact of certain historians—or even of the editors of certain anthologies. (And, of course, in identifying this tradition as "classical," I exhibit the Anglo-Saxon bias that my education was designed to inculcate: Bergson, Heidegger, and Sartre are not going to appear in *my* list of the members of anything called "the classical tradition.")

However this may be, I speak as a member of this tradition, and I want to begin by describing its presuppositions—*my* presuppositions.

According to "the classical tradition," the history of the problem of free will and determinism is, primarily, the story of a debate between two schools of philosophers, the "compatibilists" and the "incompatibilists"; that is between those who hold that free will is compatible with determinism and those who hold that free will is incompatible with determinism. Now I am going to have almost nothing to say about determinism in this paper. In fact, I am not going to talk about the problem of free will and determinism—or not directly about it. I begin with a brief characterization of the history of this problem because, while the paper is not about the problem of free

will and determinism, it presupposes the correctness of a certain way of looking at that problem. I do not propose to defend that way of looking at the problem—the way adherence to which defines membership in what I am pleased to call "the classical tradition"—, but I do want to make it clear what that way of looking at the problem is, and that it is my way. Since I shall have almost nothing to say about determinism, I shall not attempt to give any very careful explanation of this important idea. I will say only this. Determinism is the thesis that the past and the laws of nature together *determine* a unique future, that only one future is consistent with the past and the laws of nature. I am, however, going to have a great deal to say about free will and I will lay out in some detail the concept that the classical or orthodox tradition associates with the words 'free will'.

The term 'free will' is a philosophical term of art. (It is true that this term occurs in ordinary English, but its occurrence is pretty much restricted to the phrase 'of his own free will'—which means, more or less, 'uncoerced'. If someone uses the words 'free will' and does not use them within this phrase, he is almost certainly a participant in a philosophical discussion.) The first thing to realize about the use of the words 'free will' by philosophers belonging to the classical tradition is that, *now* at least, these words are a mere label for a certain feature, or alleged feature, of human beings and other rational agents, a label whose sense is not determined by the meanings of the individual words 'free' and 'will'. In particular, the ascription of "free will" to an agent by a current representative of the classical tradition does not imply that the agent has a "faculty" called 'the will'. It was not always so. Once upon a time, to say that X "had free will" was to imply that X had something called a 'will' and that this will was not only unimpeded by external circumstances (in which case the agent X *himself* was called 'free'), but that X's internal constitution left him "free" to "will" in various alternative ways. (The title of this paper is a relic of those times.) A tradition, however, is a changing thing, and the classical tradition has abandoned these implications of the words "free will." When a *current* representative of the classical tradition says of, e.g., Mrs. Thatcher, that she "has free will," he means that she is at least sometimes in the following situation: She is contemplating incompatible courses of action A and B (lecturing the Queen and holding her tongue, say), and she *can* pursue the course of action A and *can also* pursue the course of action B.

Now the word 'can' is one of the trickiest of all the little philosophically interesting Anglo-Saxon words. It is not only ambiguous; it is ambiguous in a rather complicated way. Accordingly, representatives of the classical tradition, when they are explaining the sense of their term of art 'free will', generally prefer to use some other words, in addition to 'can', to get their point across, rather as if they were trying to convey what someone looked like by displaying a photograph *and* a painted portrait *and* a pen-and-ink caricature. They say not only 'can do A and can also do B', but 'is able to do A and is also able to do B', and 'has it within his power to do A and has it within his power to do B', and 'has a choice about whether to do A or to do B'. They may also use language that is not ordinary English at all, but which seems somehow useful in conveying the sense they intend. They may, for example, talk of a sheaf of alternative possible futures that confront the agent, and say that he has free will just in the case that more than one of these futures is "open" to him or "accessible" to him.

Compatibilists, then, say that "free will" in this sense can exist in a deterministic world, and incompatibilists say that it cannot. The classical tradition sees the problem of free will and determinism as centered round the debate between the compatibilists and the incompatibilists. But what is at stake in this debate? Why should anyone care whether we have free will (in this special sense)? The answer is this: We care about morality, or many of us do, and, according to the classical tradition, there is an intimate connection between "free will" and morality. The connection is complicated, and various representatives of the classical tradition would describe it differently. But the following statement would, I think, be accepted by everyone within the classical tradition. Most within the traditional would want to say more; some *much* more. But this "highest common factor" by itself explains why many people care about whether we have free will.

> Some states of affairs are bad. They ought not to exist. And among these bad states of affairs are some that *are the fault of* certain human beings. These human beings are *to be blamed* for those states of affairs. The Nazis, for example, are to be blamed for the death camps: the existence of those camps is *their fault*. The Kennedy and Johnson and Nixon administrations are to be blamed for the U.S. involvement and actions in Vietnam. They (and perhaps others, but they

> at least) can be *held to account* for that involvement and many of its consequences. On a more homely and personal level, our profession is to blame for the fact that many young men and women are being graduated from universities who cannot compose an English sentence or tell you who Galileo was. And, doubtless, each reader of this paper knows of bad states of affairs that are his fault and his alone. But if there were no free will—if no one were able to act otherwise—then no state of affairs would be anyone's fault. No one would ever be morally accountable for anything. The actions of some people might indeed be among the causes of various bad states of affairs, but those things they caused would never be their fault. For example, suppose a father has raped his nine-year-old daughter and, as a result, she has suffered immediate physical pain and terror and has experienced life-long psychological and emotional disorders. Unless the father had at least some measure of free will, the pain and terror and the rest are not his fault. He cannot be blamed for them. They are not something for which he can be held to account.

I have not argued for this position. I am only reminding you of what the classical tradition says about the relationship between being able to do otherwise than one does and moral accountability. It is because, rightly or wrongly, the members of the classical tradition believe in this relationship that they think it is an important question whether we have free will. Almost all of the members of the classical tradition have in fact believed in free will, although there are exceptions. Baron d'Holbach believed that determinism was true and that free will was incompatible with determinism and that there was thus no free will. C.D. Broad believed that free will was incompatible with both determinism and indeterminism, and was thus impossible. But d'Holbach and Broad were exceptions. Almost all of the members of the classical tradition believe in free will. What they differ about is what free will *is*—that is, about what it is to be able to do otherwise. Most incompatibilists, at least among trained philosophers, believe in free will. All compatibilists I am aware of believe in free will; there's not much point in being a compatibilist and not believing in free will.

Before going further, I want to point out what seems to me to be

a blunder made by some writers on the problem of free will and determinism. Some writers speak of an "incompatibilist sense of 'can do otherwise'" and a "compatibilist sense of 'can do otherwise'." But when English-speaking compatibilists and incompatibilists argue about whether people could act otherwise in a deterministic world, they are using the words 'could act otherwise' in exactly the same sense. Otherwise they wouldn't be disagreeing about anything, would they? Each of them, being a speaker of English, knows what 'could have', 'was able to', and so on, mean when they are used in everyday life, and each means to be, and is, using these words in that everyday sense. Their case may be compared with the case of the dualist and the materialist in the philosophy of mind. Each uses phrases like 'feels pain' and 'is thinking about Vienna' in the same *sense*—the sense provided by the English language—though the two of them have radically opposed opinions as to the nature of the events and processes to which these terms apply. Similarly with the compatibilist and the incompatibilist: the two of them use phrases like 'could have acted otherwise' in just the same sense—the sense provided by the English language—and disagree about whether that one sense expresses something that could obtain in a deterministic world. Now it may be that a particular compatibilist or incompatibilist has a mistaken *theory* about what 'could act otherwise' means. But, in such a case, that philosopher does not *himself* mean by 'could act otherwise' what his mistaken theory says these words mean. For example, suppose that a certain compatibilist has published an essay the burden of which is that '*x* could act otherwise' means '*x* would act otherwise if he chose to'. And suppose that this is wrong: suppose that this is not a correct account of the meaning of the English phrase 'could act otherwise'. Then that compatibilist is not only wrong about what others mean by 'could act otherwise'; he is also wrong about what *he* means by these words. (Compare this case: if I mistakenly think that 'knowledge' means 'justified true belief', it does not follow that that is what I mean by 'knowledge'.) If philosophers always used words to mean what their theories said those words meant, no philosopher would ever revise a definition because of a counter-example. But this occasionally happens. Now if all anyone means by talk of an "incompatibilist sense" or a "compatibilist sense" of the central terms in the free-will debate is that philosophers have sometimes proposed theories about the meanings of these terms, theories that support compatibilism (or, it may be, incompatibilism),

I have no objection. But then we must remember that it remains an open question whether compatibilists use these terms in a "compatibilist sense" and whether incompatibilists use these terms in an "incompatibilist sense."

Finally, it is this single sense of 'can do otherwise', the sense provided by ordinary English, that compatibilists and incompatibilists contend is so intimately connected with the possibility of moral accountability. This is the classical tradition.

Let me now turn to my title. My question is, just how often is it that we are able to do otherwise? A belief in one's free will is the belief that one can sometimes do otherwise. But then it is consistent to say of X that he has free will despite the fact that he can almost never do otherwise. The central thesis of this paper is that while it is open to the compatibilist to say that human beings are very often—hundreds of times every day—able to do otherwise, the incompatibilist must hold that being able to do otherwise is a comparatively rare condition, even a *very* rare condition.

It is almost self-evident that compatiblism entails that being able to do otherwise is as common as pins. Or, at any rate, it is evident that typical versions of compatibilism entail this. Typical versions of compatibilism entail that being able to do otherwise is some sort of conditional causal power. For example, one primitive version of compatibilism—a version pretty generally agreed to be unsatisfactory— holds that for one to have been able to act differently is for one to have been such that one would have acted differently if one had chosen to act differently. (More generally, for one to be able to do A is for one to be such that one would do A if one chose to.) And who could deny that at most moments each of us is such that he would then be acting differently if he had chosen to act differently?

The case is otherwise with incompatibilism. To see why this is so, let us remind ourselves of why people become incompatibilists. They become incompatibilists because they are convinced by a certain sort of argument. My favorite version of it—which I reproduce from my book *An Essay on Free Will*[1]—turns on the notion of "having a choice about." Let us use the operator 'N' in this way: 'N*p*' stands for '*p* and no one[2] has, or ever had, any choice about whether *p*'. The validity of the argument turns on the validity of two rules of deduction involving 'N':

Rule Alpha: From $\Box p$ deduce Np. ('$\Box$' represents "standard necessity": truth in all possible circumstances.)
Rule Beta: From Np and $N(p \supset q)$ deduce Nq.

Now let 'P' represent any true proposition whatever. Let 'L' represent the conjunction into a single proposition of all laws of nature. Let 'P_0' represent a proposition that gives a complete and correct description of the whole world at some instant in the remote past—before there were any human beings. If determinism is true, then $\Box(P_0 \,\&\, L. \supset P)$. We argue from this consequence of determinism as follows.

1. $\Box(P_0 \,\&\, L. \supset P)$	
2. $\Box(P_0 \supset (L \supset P))$	1; modal and sentential logic
3. $N(P_0 \supset (L \supset P))$	2; Rule Alpha
4. NP_0	Premise
5. $N(L \supset P)$	3, 4; Rule Beta
6. NL	Premise
7. NP	5, 6; Rule Beta

If this argument is sound, then determinism entails that no one has or ever had any choice about anything. Since one part of "anything" is what any given person does, this amounts to saying that determinism entails that no one could ever have done otherwise. No one, I think, could dispute the two premises or Rule Alpha. The question of the soundness of the argument thus comes down to the question whether Rule Beta is valid. It is not my purpose in this paper to defend Beta. I reproduce this argument only to point out the central role that Beta (or something equivalent to it) plays in the incompatibilist's reasons for accepting his theory. I will go so far as to say that, in my view, one could have no reason for being an incompatibilist if one did not accept Beta. If one accepts Beta, one should be an incompatibilist, and if one is an incompatibilist, one should accept Beta.

What I propose to show in the sequel is this: Anyone who accepts Beta should concede that one has precious little free will, that rarely, if ever, is anyone able to do otherwise than he in fact does. I shall argue for this position as follows. I shall first show that if Rule Beta is valid, then no one is able to perform an act he considers morally reprehensible. I shall then extend this argument; by a similar sort of reasoning, I shall show that, given Rule Beta, no one is able to

do anything if he wants very much *not* to do that thing and has no countervailing desire to do it. Finally, by more or less the same reasoning, I shall show that the validity of Rule Beta entails that if we regard an act as the one obvious thing or the only sensible thing to do, we cannot do anything but that thing.

In *Elbow Room*[3], Daniel Dennett has argued eloquently that he is simply *unable* to do anything he regards as morally reprehensible. Compatibilists may feel a bit uneasy about agreeing with Dennett about this. Really simple-minded and primitive compatibilists, those who hold that one can do something just in the case that one would do it if one chose to, *must* disagree with Dennett. Take Dennett's primary example, the torture of an innocent victim in return for a small sum. Dennett will concede, I am sure, that we can easily imagine situations in which he, being more or less as he is now, would succeed in carrying out such torture *if he chose to*. His point is that, being as he is, he would never choose to. *I* think that this is a perfectly good point, but, of course, it is a point that must be disallowed by the primitive compatibilist who identifies the ability to perform an act with the absence of environmental impediments to performing that act. Leaving aside the question of what more sophisticated compatibilists might say about such cases, let us turn to the incompatibilists. They, I maintain, must agree with Dennett. Dennett uses himself as an example. I will use myself. Let us consider some act I regard as reprehensible. I might, like Dennett, use torture as an example, but my acquaintance with torture is purely literary, and I should like to try to avoid that dreamlike sense of unreality that is so common in philosophical writing about morality. I will pick an example that touches my own experience. Recently, a member of my university, speaking on the floor of a College meeting, deliberately misrepresented the content of the scholarly work of a philosopher (who was not present), in an attempt to turn the audience against him. Suppose such a course of action were proposed to me. Suppose someone were to say to me, "Look, you don't want Smith to be appointed Chairman of the Tenure Committee, so tell everyone that he said in print that all sociologists are academic charlatans. (I've got a quotation you can use that seems to say that if you take it out of context.) Then the sociologists will block the appointment." Call the act that is proposed A. I regard lying about someone's scholarly work as reprehensible. And, while I should prefer not to see Smith appointed, I certainly wouldn't *think* of blocking

his appointment by any such means. In short, I regard the proposed act A as being indefensible. (I mean in the actual circumstances: I might lie about the content of someone's scholarly work to prevent World War III, but the start of World War III is not *in fact* what hangs on my performing or not performing A.) I may even say that I regard doing A as being "indefensible, given the totality of information available to me." And, of course, I do not so regard *not* doing A: there's nothing much to be said against that. We may also suppose that I am unable (as things stand) to search out any further relevant information—the vote will come in a moment, and I must speak at once if my speaking is to affect it. Now consider the following conditional:

> *C* If X regards A as an indefensible act, given the totality of relevant information available to him, and if he has no way of getting further relevant information, and if he lacks any positive desire to do A, and if he sees no objection to *not* doing A (again, given the totality of relevant information available to him),
> *then* X is not going to do A.

What is the modal status of *C*? It seems to me to be something very like a necessary truth. What would be a conceivable circumstance in which its antecedent is true and its consequent false (i.e., X proceeds to do A)? If X changes his mind about the indefensibility of A (perhaps because of the intervention of some "outside" agent or force, or because of an access of new information, or because he suddenly sees some unanticipated implication of the information available to him)? If X just goes berserk? If so, build the non-occurrence of these things into the antecedent of *C*: he is not going to change his mind about the indefensibility of A and he is not going to go berserk.

It seems to me that there is no possible world in which C is false. What would it be *like* for *C* to be false? Imagine that X *does* do A. We ask him, "Why did you do A? I thought you said a moment ago that doing A would be reprehensible." He replies:

> Yes. I did think that. I still think it. I thought that at every moment up to the time at which I performed A; I thought that while I was performing A; I thought it immediately afterward. I never wavered in my conviction that A was an irremediably reprehensible act. I never

> thought there was the least excuse for doing A. And don't misunderstand me: I am not reporting a conflict between duty and inclination. I didn't *want* to do A. I never had the least desire to do A. And don't understand me as saying that my limbs and vocal cords suddenly began to obey some will other than my own. It was *my* will that they obeyed. It is true without qualification that *I* did A, and it is true without qualification that I *did* A.

This strikes me as absolutely impossible. It's not, of course, impossible for someone to say these words—just as it's not impossible for someone to say, "I've just drawn a round square." But it is impossible for someone to say these words and thereby say something true.

Now consider the proposition that I consider the act A to be indefensible. I think it's pretty clear that I have—right now—no choice about how I feel about A. Like most of my beliefs and attitudes, it's something I just find myself with. (Which is not to say that I don't think that this attitude is well-grounded, appropriate to its object, and so on.) If you offered me a large sum of money, or if you promised—and I believed you could deliver—the abolition of war, if only I were to change my attitude toward A, I should not be able to take you up on this offer, however much I might want to. It is barely conceivable that I have the ability to change my attitude toward A over some considerable stretch of time, but we're not talking about some considerable stretch of time; we're talking about right now.

Let us now examine a certain Beta-like rule of inference, which I shall call Beta-prime:

> From N x,p and N x,($p \supset q$) deduce N x,q.

Here 'N' is a two-place operator, and 'N x,p' abbreviates 'p and x now has no choice about whether p'. The one-place operator 'N' served my purposes in *An Essay on Free Will*, because there the premises of my argument concerned only propositions that were related in just the same way to all human beings, past, present, and future: laws of nature and propositions about the state of the world before there were any human beings. It is clear, I think, that whatever relation any given human being bears to such a proposition, any other given human being bears that relation, too. Since I was interested only in such propositions, I employed the impersonal and timeless

one-place 'N'; it was simpler to do so. The arguments I wish to consider in the present paper, however, involve propositions about particular human beings and what they do at particular times, and their attitudes toward what they do at those times. For that reason, I need to use the person- and time-relative rule Beta-prime, and I must forego the convenience of Beta. And Beta-prime seems hardly less evident than Beta. The same intuitive considerations that support Beta seem to support Beta-prime, and it is hard to imagine a philosopher who accepts Beta but rejects Beta-prime.

Consider the following instance of Beta-prime:

N I, I regard A as indefensible
N I, (I regard A as indefensible ⊃ I am not going to do A)
hence, N I, I am not going to do A.

In this argument, 'I regard A as indefensible' is short for 'I regard A as an indefensible act, given the totality of relevant information available to me, and I have no way of getting further relevant information, and I lack any positive desire to do A, and I see no objection to *not* doing A, given the totality of relevant information available to me.' (Compare the antecedent of the conditional *C*, above.) The conclusion of this argument, written out in full, is 'I am not going to do A and I now have no choice about whether I am not going to do A'. Now the second conjunct of this sentence is a bit puzzling. But we may note that the sentences 'I have a [or *no*] choice about whether *p*' and 'I have a [or *no*] choice about whether not-*p*' would seem to be equivalent. Therefore, we may read the conclusion of the argument as 'I am not going to do A and I now have no choice about whether I am going to do A'. (The reason the original version of the conclusion seems puzzling is this: the mind looks for a function for that final 'not' to perform and finds none.)

The first premise of this argument is true, because, as we have seen, I (right now, at any rate) have no choice about whether I regard A as indefensible. The second premise is true because as we have seen, the conditional 'I regard A as indefensible ⊃ I am not going to do A' is a necessary truth, and no one has any choice about the truth-value of a necessary truth.

The general lesson is: if I regard a certain act as indefensible, then it follows not only that I *shall not* perform that act but that I *can't* perform it. (Presumably, 'I am not going to do A and I have no choice about whether I am going to do A' is equivalent to 'I can't do A'.)

This conclusion is not intuitively implausible. To say that you can do A (are able to do A, have it within your power to do A) is to say something like this: there is a sheaf of alternative futures spread out before you; in some of those futures you do A; and some at least of those futures in which you do A are "open" to you or "accessible" to you. Now if this picture makes sense (as a picture; it's only a picture), it would seem to make sense to ask what these futures are like. You say you can do A; well, what would it be like if you did? You say that a future in which you do A is "open" to you or "accessible" to you? Well, in what circumstances would you find yourself if you "got into" or "gained access to" such a future? If you can't give a coherent answer to this question, that, surely, would cast considerable doubt on your claim to be able to do A.

And suppose I do regard doing A as indefensible (for me, here, now). Then, I think, I cannot give a coherent description of a future (one coherently connected with the present) in which I proceed to do A. I have already considered what such an attempt would sound like ("Yes. I did think that...") and have rejected it—rightly—as incoherent.

We must conclude, therefore, that (given the validity of Beta-prime) I *cannot* perform an act I regard as indefensible, and that this is a perfectly intuitive thesis. Its connection with incompatibilism is displayed in the following argument.

(1) If the rule Beta-prime is valid, I cannot perform an act I regard as indefensible.
(2) If the rule Beta is valid, the rule Beta-prime is valid.
(3) Free will is incompatible with determinism only if Beta is valid.

hence,

(4) If free will is incompatible with determinism, then I cannot perform an act I regard as indefensible.

Throughout this little argument, 'I cannot perform an act I regard as indefensible' is to be understood in a *de re*, not a *de dicto* sense. It does not mean, 'Not possible: I perform an act I regard as indefensible'; it means, 'For any act x, if I regard x as indefensible, then I do not have it within my power to perform x'. (I don't mean to deny the *de dicto* statement; it is in fact true, but it doesn't figure in the argument.)

The defense of premise (1) of this argument has been the main task of the paragraphs preceding the argument.

Premise (2) seems undeniable because, as I have said, the intuitions that support Beta also support Beta-prime.

Premise (3) can be defended on this ground: the only reason known for accepting incompatibilism is that it follows from Beta. This, of course, does not *prove* that (3) is true. But it is unlikely that anyone would accept incompatibilism and reject Beta.

Let us now leave the topic of indefensibility and turn to desire—to cases of simple, personal desire having no moral dimension whatever.

Suppose that someone has an (occurrent) desire to perform some act. Suppose that this desire is very, very strong, and that he has no countervailing desire of any sort. (We have considered the case in which duty is unopposed by inclination. We now turn to the case in which inclination is unopposed by inclination.) Consider the case of poor Nightingale in C.P. Snow's novel *The Masters*. Nightingale wants to be a Fellow of the Royal Society—in the idiom of the 1980s, he wants this distinction so badly he can taste it. Every year, on the Royal Society's election day, Nightingale strides out to the porter's lodge of his Cambridge college and leaves *strictest* instructions that, if a telegram arrives for him, he is to be notified *immediately*. (He threatens the porter with summary dismissal if there is the slightest delay.) Now suppose that poor Nightingale, on the day of the election, is sitting in his rooms biting his nails and daydreaming about being able to call himself 'F.R.S.'. The telephone rings. He snatches it from its cradle and bawls, "Nightingale here," doubtless deafening his caller.

What I want to know is: *Could* he have refrained from answering the telephone? Was he able not to touch it? Did he have it within his power to let it ring till it fell silent? If what we have said above (in connection with indefensibility) is correct, he could have refrained from answering the telephone only if we can tell a coherent story (identical with the story we *have* told up to the point at which the telephone rings) in which he *does* refrain from answering the telephone. Can we? Well, we might tell a story in which, just as the telephone rings, Nightingale undergoes a sudden religious conversion, like Saul on the road to Damascus: All in a moment, his most fundamental values are transformed and he suddenly sees the Fellows of the Royal Society as cocks crowing on a dunghill. Or we might imagine that Nightingale's mind snaps at the moment the telephone rings and he begins to scream and break up furniture and eventually has to be put away. But, remember, neither of these things *did*

happen. Let's suppose that they did not even come close to happening. Let's suppose that there was at the moment we are considering no disposition in the mind of God or in Nightingale's psyche (or wherever the impetus to religious conversion is lodged) toward a sudden change in Nightingale's most fundamental values. Let's suppose also that the moment at which the telephone rang was the only moment at which there was *no* possibility of Nightingale's mind snapping—it was a moment of sudden, intense hope, after all. Build these suppositions into our story of how it was with Nightingale up till the moment at which the telephone rang. Build into it also the proposition that no bullet or lightning bolt or heart attack is about to strike Nightingale. Call this story the Telephone Story. I am inclined to think that there is no possible world in which the Telephone Story is true and in which Nightingale does not proceed to answer the telephone. We have the following instance of the rule Beta-prime (imagine that the present moment is the moment at which the telephone rings):

> N Nightingale, the Telephone Story is true.
> N Nightingale, (the Telephone Story is true ⊃ Nightingale is going to answer the telephone)
> *hence*,
> N Nightingale, Nightingale is going to answer the telephone.

The conclusion may be paraphrased, 'Nightingale is going to answer the telephone, and he has no choice about whether to answer the telephone'. And the premises seem undeniably true.

The lesson would seem to be: If the rule Beta-prime is valid, then if a person has done A, and if he wanted very much to do A, and if he had no desires whatever that inclined him towards not doing A, then he was unable not to do A; not doing A was simply not within his power. An argument similar to the one given above shows that the incompatibilist ought to accept this consequence of Beta-prime.

Let us, finally, turn to a third kind of case. On many occasions in life, with little or no deliberation or reflection, we simply do things. We are not, on those occasions, in the grip of some powerful desire, like poor Nightingale. The things just seem—or would seem if we reflected on them at all—to be the obvious things to do in the circumstances. I suppose that on almost all occasions when I have answered the telephone, I have been in more or less this position. On most occasions on which I have answered the telephone, I have

not been biting my nails in a passion of anxiety and impatience like Nightingale. On most such occasions, I have not been expecting the telephone to ring (not that its ringing *violates* any expectation of mine, either); with my mind still half on something else, I pick up the receiver and absently say, "Hello?" Obviously, mere habit has a lot to do with this action, but I do not propose to inquire into the nature of habit or into the extent of its involvement in such acts.

Now consider any such occasion on which I answered the telephone. I was sitting at my desk marking papers (say); the telephone rang. (I had not been expecting it to ring. I had no reason to suppose it would *not* ring.) I answered it. Without reflection or deliberation. I simply put down my pen and picked up the receiver.

Can we tell a coherent story in which (in just those circumstances) I simply ignore the telephone and go on marking papers till it stops ringing? Well, we might. Since the matter is a minor one, we need not postulate anything on the order of a religious conversion. We might simply assume that some good reason for not answering the telephone suddenly popped into my mind. (Didn't I have a letter recently from a man who claimed to be able to prove mind-body dualism from the fact that he had made several trips to Mercury by astral projection? Didn't he say that he would be calling me today to make an appointment to discuss the implications of his astral journey for the mind-body problem?) Or, again, we might imagine that I suddenly go berserk and begin to smash furniture. Or we might postulate a sudden Divine or meteorological or ballistic alteration of my circumstances. But we might also imagine that there exists no basis either in my psyche or my environment (at the moment the telephone rings) for any of these things. We may even, if you like, suppose that at the moment the telephone rings it is causally determined that no reason for not answering the phone will pop into my mind in the next few seconds, and that it is causally determined that I shall not go berserk or be struck dead.

This set of statements about me and my situation at the moment the telephone rang (and during the two or three minutes preceding its ringing) we may call the Second Telephone Story. It seems to me to be incoherent to suppose that the Second Telephone Story is true and that I, nevertheless, do not proceed to answer the telephone. And, of course, we have the following instance of Beta-prime:

N I, The Second Telephone Story is true.
N I, (The Second Telephone Story is true ⊃ I am going to answer the telephone).
hence,
N I, I am going to answer the telephone.

The conclusion may be read: 'I am going to answer the telephone and I have no choice about whether to answer the telephone'. Its connection with incompatibilism can be established by an argument not essentially different from the one already given.

It seems clear that if the premises of this third instance of Rule Beta-prime are true, then we have precious little free will—at least assuming that Beta-prime is valid. For our normal, everyday situation is represented in the Second Telephone Story. It is perhaps not clear how many of the occasions of everyday life count as "making a choice." The light turns green, and the driver, his higher faculties wholly given over to thoughts of revenge or lunch or the Chinese Remainder Theorem puts his car into gear and proceeds with his journey. Did he do something called "making a choice between proceeding and not proceeding"? Presumably not: the whole thing was too automatic. The young public official, unexpectedly and for the first time, is offered a bribe, more money than he has ever thought of having, in return for an unambiguous betrayal of the public trust. After sweating for thirty seconds, he takes the money. Did he make a choice? Of course. Between these two extremes lie all sorts of cases, and it is probably not possible to draw a sharp line between making a choice and acting automatically. But I think it is evident that, wherever we draw the line, we are rarely in a situation in which the need to make a choice confronts us and in which it isn't absolutely clear what choice to make. And this is particularly evident if we count as cases of its being "absolutely clear what choice to make" cases on which it is absolutely clear *on reflection* what choice to make. A man may be seriously considering accepting a bribe until he realizes (after a moment's reflection on the purely factual aspects of his situation) that he couldn't possibly get away with it. Then his course is clear, because it has become clear to him that there is nothing whatever to be said for taking the bribe and a great deal to be said against it. He has not *decided* which of two incompatible objects of desire (riches and self-respect, say) to accept; rather he has *seen* that one of the two—riches—wasn't really there.

There are, therefore, few occasions in life on which—at least after a little reflection and perhaps some investigation into the facts—it isn't absolutely clear what to do. And if the above arguments are correct, then an incompatibilist should believe that on such occasions the agent cannot do anything other than the thing that seems to him to be clearly the only sensible thing.

Now there are *some* occasions on which an agent is confronted with alternatives and it is not clear to him what to do—not even when all the facts are in, as we might put it. What are these cases like? I think we may distinguish three cases.

First, there are what might be called "Buridan's Ass" cases. Someone wants each of two or more incompatible things and it isn't clear which one he should (try to) get, and the things are interchangeable; indeed their very interchangeability is the reason why it isn't clear to him which to try to get. (I include under this heading cases in which the alternatives are importantly different but look indistinguishable to the agent because he unavoidably lacks some relevant datum. Lady-and-tiger cases, we might call them.) Closely allied with Buridan's Ass cases, so closely that I shall not count them constituting a different kind of case, are cases in which the alternatives are not really interchangeable (as are two identical and equally accessible piles of hay) but in which the properties of the alternatives that constitute the whole of the difference between them are precisely the objects of the conflicting desires. We might call such cases "vanilla/chocolate cases." They are often signaled by the use of the rather odd phrase 'I'm trying to decide which one I want'—as opposed to '...which one to have'. I want chocolate and I want vanilla and I can't (or won't or don't want to) have both, and there is no material for deliberation, because my choice will have no consequences beyond my getting vanilla, or, as the case may be, chocolate. (Note, by the way, that someone who is trying to decide whether to have chocolate, to which he sometimes has an allergic reaction, or vanilla, which he likes rather less than chocolate, does not constitute what I am calling a "vanilla/ chocolate case.") Both vanilla/chocolate cases and "Buridan's Ass proper" cases are characterized by simple vacillation. Hobbes's theory of deliberation, whether or not it is satisfactory as a general theory, is pretty uncontroversially correct in these cases. One wavers between the alternatives until one inclination somehow gets the upper hand, and one ends up with a chocolate cone or the bale of hay on the left.

The second class of cases in which it is not obvious what to do (even when all the facts are in) are cases of duty versus inclination. Or, better, cases of general policy versus momentary desire. (For what is in conflict with the agent's momentary desire in such cases need have nothing to do with the agent's perception of his moral duty; it might have no higher object than his long-term self-interest.) I have made for myself a maxim of conduct, and no sooner have I done this than, in St. Paul's words, "...I see another law in my members, warring against the law of my mind." Our story of the young official and the proffered bribe is an example; further examples could be provided by any dieter. This class of cases is characterized by what is sometimes called moral struggle, although, as I have said, not all cases of it involve morality.

The third class of cases involves incommensurable values. (I owe this point to the work of Robert Kane.[4]) A life of rational self-interest (where self-interest is understood to comprise only such ends as food, health, safety, sex, power, money, military glory, and scientific knowledge, and not ends like honor, charity, and decency) versus a life of gift and sacrifice; caring for one's aged mother versus joining the Resistance; popularity with the public versus popularity with the critics. All these are cases of incommensurable values. Other cases would have to be described with more care to make sure that they fit into this class. The case of a young person wondering whether to become a lawyer or a concert pianist might belong to this class. But not if the question were, "In which profession should I make more money?", or "In which profession should I make the greater contribution to human happiness?" In those cases, values are not at issue, but only how maximize certain "given" values; the matter is one of (at best) calculation and (at worst) guesswork. The general form of the question that confronts the agent in true cases of the third type is, What sort of human being shall I be?, or What sort of life shall I live? And, of course, this does not mean, What sort of life is dictated for me by such-and-such values (which I already accept)? *That* question is one to be decided by calculation or guesswork. In cases of the third type, the agent's *present* system of values does not have anything to tell him. His values may tell him to become a professional rather than a laborer and an honest rather than a dishonest professional, but they do not tell him whether to become a lawyer or a pianist. (It may be that the values he could expect to have as a result of the choice would confirm that choice—see Kierkegaard

on the moral versus the "aesthetic" life—but that's of no help to him *now*.) The choices in the third category are those that many philosophers call "existential," but I will not use this term, which derives from a truly hopeless metaphysic. As the cases in the first category are characterized by vacillation, and the cases in the second by "moral" struggle, so the cases in the third are characterized by *indecision*—often agonized indecision. The period of indecision, moreover, may be a long one: weeks, months, or even a really significant part of the agent's life.

I believe that these three cases exhaust the types of case in which it is not obvious to the agent, even on reflection, and when all the facts are in, how he ought to choose. Therefore, if our previous arguments are correct, the incompatibilist should believe that we are faced with a genuinely free choice only in such cases. (That is: in these cases, if in any. The incompatibilist may well believe that in some of *these* cases we have no choice about how to act, or, like d'Holbach and C.D. Broad, that even in these cases we have no choice about how to act.) It is not clear to *me* that in cases of the first type—"Buridan's Ass" cases—there is any conceivable basis for saying that we have a choice about what to do. Doubtless when we choose between identical objects symmetrically related to us, or when we choose between objects that differ only in those properties that are the objects of our competing desires, there occurs something like an internal coin-toss. (My guess, for what it's worth, is that we contain a "default" decision-maker, a mechanism that is always "trying" to make decisions—they would be wholly arbitrary decisions if it were allowed to make them—but which is normally overridden by the person; I speculate that when "vacillation" occurs, the person's control over the "default" decision maker is eventually suspended and it is allowed to have its arbitrary way.) I think that it's pretty clear that in such cases one has no choice about how one acts. If one tosses a coin, then one has no choice about whether it will land heads or tails. And, indeed, why should one want such a power—if the alternatives really are indifferent?

If this is correct, then there are at most *two* sorts of occasion on which the incompatibilist can admit that we exercise free will: cases of an actual struggle between perceived moral duty or long-term self-interest, on the one hand, and immediate desire, on the other; and cases of a conflict of incommensurable values.

Both of these sorts of occasion together must account for a fairly

small percentage of the things we do. And, I must repeat, my conclusion is that this is the *largest* class of actions with respect to which the incompatibilist can say we are free. The argument I have given shows that the incompatibilist ought to deny that we have free will on any occasions other than these. It has no tendency to show that the incompatibilist should say that we do act freely on these occasions. The argument purports to show that, given the principles from which the incompatibilist derives his position, it is impossible for us to act freely on occasions other than these. It has no tendency to show that—given the incompatibilist's principles—it is possible for us to act freely on any occasion whatever. It's like this: A biologist, using as premises certain essential features of mammals and some facts about Mars, proves that there could not be mammalian life on Mars; such a proof, even if it is beyond criticism, has no tendency to show that there *could* be any sort of life on Mars. That's as may be. His proof just tells us nothing about non-mammalian life.

I will not discuss (further) the question of how much free will we might have *within* these two categories. In the sequel, I wish to discuss the implications of what I have argued for so far for questions of moral blame.

I have argued that, if incompatibilism is true, free action is a less common phenomenon than one might have thought. It does not, however, follow that moral accountability is a less common phenomenon than one might have thought. And this is the case even on the traditional or "classical" understanding of the relationship between free will—that is, the power or ability to do otherwise than one in fact does—and accountability. Nothing that has been said so far need force the incompatibilist (the incompatibilist whose view of the relation between free will and blame is that of the classical tradition) to think that moral accountability is uncommon.

Let us see why. Would anyone want to say that the classical tradition is committed to the following thesis? "An agent can be held accountable for a certain state of affairs only if either (a) that agent intentionally brought that state of affairs about and could have refrained from bringing it about, or (b) that agent foresaw that that state of affairs would obtain unless he prevented it, and he was able to prevent it." I don't know whether anyone would want to say this. My uncertainty is due mainly to the fact that philosophers discussing problems in this general area usually talk not about accountability for states of affairs— the *results* of our action and inaction—but ac-

countability (or "responsibility") for *acts*. This way of talking about these matters is confusing and tends to obscure what I regard as crucial points. However this may be, the classical tradition is not committed to this thesis, though it may be that some representatives of the tradition have endorsed it. This is fortunate for the tradition, because the thesis is obviously false. This is illustrated by "drunk driver" cases: I could not have swerved fast enough to avoid hitting the taxi, and yet no one doubts that I am to blame for the collision. How can that be? Simple: I was drunk and my reflexes were impaired. Although I was unable to swerve to avoid hitting the taxi, that inability (unlike, say, my inability to read minds) was one I could have avoided having. Or again: Suppose that when I am drunk it is not within my power to refrain from violently assaulting those who disagree with me about politics. I get drunk and overhear a remark about Cuban troops in Angola and, soon thereafter, Fred's nose is broken. I was, under the circumstances, unable to refrain from breaking Fred's nose. And yet no one doubts that I am to blame for his broken nose. How can that be? Simple: Although I was unable to avoid breaking his nose, that inability is one I could have avoided having. What these examples show is that the inability to prevent or to refrain from causing a state of affairs does not logically preclude being to blame for that state of affairs. Even the most orthodox partisan of a close connection between free will and blame will want to express this connection in a principle that is qualified in something like the following way:

> An agent cannot be blamed for a state of affairs unless there was a time at which he could so have arranged matters that that state of affairs not obtain.

And this principle is at least consistent with its being the case that, while we are hardly ever able to act otherwise than we do, we are nevertheless accountable for (some of) the consequences of *all* of our acts. (No one, I suppose, would seriously maintain that we can be blamed for *all* of the consequences of *any* of our acts. If I am dilatory about returning a book to the library and this has the consequence—apparent, I suppose, only to God—that a certain important medical discovery is never made, the thousands of deaths that would not have occurred if I had been a bit more conscientious are not my fault. And who can say what the unknown consequences of our most casual acts may be? Obviously, I can be blamed only for those consequences of my acts that are in some sense "foreseeable.") Consider this case.

A Mafia hit-man is dispatched to kill a peculating minor functionary of that organization. The victim pleads for his life in a most pathetic way, which so amuses the hit-man (who would no more think of failing to fulfill the terms of a contract than you or I would think of extorting money from our students by threats of failing them) that he shoots the victim in the stomach, rather than through the heart, in order to prolong the entertainment. Could he have refrained from killing the victim? Was it, just before he shot the victim, within his power to pocket his gun unfired and leave? If what has been said so far is true, probably not. Would it follow that he was not morally responsible for the victim's death? By no means. Given the kind of man he was, he was unable, in that situation, to have acted otherwise. But perhaps he could have avoided having that inability by avoiding being the kind of man he was. It is an old, and very plausible, philosophical idea that, by our acts, we make ourselves into the sorts of people we eventually become. Or, at least, it is plausible to suppose that our acts are *among* the factors that determine what we eventually become. If one is now unable to behave in certain ways—I am not talking about gross physical inabilities, like a double amputee's inability to play the piano—this may be because of a long history of choices one has made. Take the case of cold-blooded murder. The folk wisdom has it (I don't know if there is empirical evidence for this) that most of us have been born with a rather deep reluctance to kill helpless and submissive fellow human beings. But, if there is such a reluctance, it can obviously be overcome. And (so the folk wisdom has it) each time this reluctance is overcome it grows weaker, until it finally disappears. Suppose our Mafia hit-man *did* have a free choice the first time he killed a defenseless victim. He might have experienced on that occasion—though doubtless these terms were not in his vocabulary— something like a conflict between momentary inclination and long-term self-interest. Suppose he did kill his man, however, and that he continued to do this when it was required of him until he had finally completely extirpated his reluctance to kill the helpless and submissive. If he is now unable to pocket his gun unfired and walk away, this is, surely, partly because he has extirpated this reluctance. The absence of this normal reluctance to kill is an essential component of his present inability not to kill. If the folk wisdom is right, and let us suppose for the sake of the example that it is, then it is conceivable he could have avoided having his present inability. And, therefore, it may be, for all we have said,

that he can properly be held to account for the victim's death. Given the causal and psychological theses contained in the folk wisdom, he may be accountable for the victim's death for the same reason that a drunk driver is accountable for an accident traceable to his impaired reflexes. (But, of course, I don't mean to suggest that the case of a man who has turned himself into a sociopath by a long series of free choices over many years is morally *very* much like the case of a man who has turned himself into a temporarily dangerous driver by one or two acts of free choice in the course of an evening.)

I have nothing more to say on the subject of moral blame. This is a difficult topic, and one that involves many other factors than the ability to act otherwise. (Coercion and ignorance, for example, are deeply involved in questions of accountability. And there is the dismally difficult question of what it is for a consequence of an act to be "foreseeable" in the relevant sense.) My only purpose in these last few paragraphs has been to give some support to the idea that the radically limited domain of the freedom of the will that the incompatibilist must accept does not obviously commit him to a similarly radically limited domain for moral blame. It may be that we are usually right when we judge that a given state of affairs is a given person's fault, even if people are almost never able to refrain from bringing about the states of affairs they intentionally bring about, and even if people are almost never able to act to prevent the states of affairs that they know perfectly well will obtain if they do not act to prevent them. For it may be that they could have avoided having these inabilities.[5]

Notes

1. See pp. 93-105.
2. That is, no human being. We shall not take into account the powers of God or angels or Martians.
3. See pp. 133ff.
4. See [2], Part II.
5. This paper was read at a conference on "Freedom and Mind" at McGill University in September, 1986, and as an invited paper at the 1987 meeting of the Central Division of the American Philosophical Association. On the latter occasion, the commentator was R. Kane. The paper was also read to the Philosophy Department at Virginia Polytechnic Institute and State University. The audiences on these occasions are

thanked for their useful comments, as are those who have been kind enough to correspond with me about the topics discussed herein. Special thanks are due to Daniel Dennett, Robert Kane, and Lawrence H. Davis.

References

1. Daniel C. Dennett, *Elbow Room: The Varieties of Free will Worth Wanting* (Cambridge, Mass. and London: Bradford Books, 1984).
2. R. Kane, *Free Will and Values* (Albany: State University of New York Press, 1985).
3. Peter van Inwagen, *An Essay on Free Will* (Oxford: The Clarendon Press, 1983).