12

The Incompatibility of Responsibility and Determinism

PETER VAN INWAGEN

Many philosophers think that determinism is incompatible with moral responsibility. Probably most of the philosophers who accept this thesis accept it on the basis of some argument very much like this one:

(i) Determinism is incompatible with free will
(ii) Moral responsibility is impossible without free will
∴ Determinism is incompatible with moral responsibility.

I am one of these philosophers.[1] I think that both (i) and (ii) are true and I believe that I am in possession of good reasons for thinking this. I am aware, however, that many philosophers think (i) is false. Many philosophers, in fact, think that anyone who accepts (i) convicts himself *ipso facto* of philosophical incompetence.[2] (I may remark that this attitude evidences very high standards of philosophical competence indeed, since among the philosophers who accept (i) are professors Anscombe, Chisholm, and Plantinga.) Because (i) is so very controversial, however, I propose in this paper to investigate the question

"The Incompatibility of Responsibility and Determinism" originally appeared in *Bowling Green Studies in Applied Philosophy*, vol. 2, ed. M. Bradie and M. Brand (Bowling Green, Ohio: Bowling Green State University, 1980), 30–37, © 1979 by Bowling Green State University, and is reprinted here with permission from the Editor.

1. See my articles "A Formal Approach to the Problem of Free Will and Determinism," *Theoria*, Vol. XL, Part 1 (1974) and "The Incompatibility of Free Will and Determinism," *Philosophical Studies* 27 (1975).

2. See, e.g., the opening paragraphs of Donald Davidson's "Freedom to Act" in Ted Honderich, ed., *Essays on Freedom of Action* (London: Routledge & Kegan Paul, 1973).

whether moral responsibility is compatible with determinism *independently* of (i). I shall argue that determinism and responsibility are incompatible, and not only shall I make no use of proposition (i) in my argument, I shall make no mention whatever of free will other than a very brief one at the end of the paper, and that in relation to a question of secondary importance. I concede that my argument will bear a certain structural resemblance to various arguments for the incompatibility of free will and determinism, but that is neither here nor there: the concept of free will will not *figure* in my argument.

I

In the remainder of this paper, I shall often drop the word 'moral' and speak simply of responsibility. But I mean my remarks about responsibility to apply only to moral responsibility. I do not claim for example, that everything I say about "responsibility" is true of legal responsibility. I shall offer no definition or analysis of responsibility. I have no analysis to give and I doubt whether an analysis of responsibility would contribute much to my argument in any case. I *shall* argue that certain propositions involving the concept of responsibility are conceptual truths, but I am pretty sure I should simply reject any proposed analysis of responsibility that was in conflict with the conceptual claims I am going to make about responsibility. For example, I shall have occasion to claim that it is a conceptual truth that no human being can be held responsible for the way the world was before there had ever been any human beings, and if someone were to propose an analysis of responsibility that had the consequence that some human being *could* be held responsible for some preadamite state of affairs, then we should have the right to be certain, without further inquiry, that his analysis was wrong. I do not mean that I shall not defend my claims about conceptual truths involving the notion of responsibility. I shall. But my defences will be informal and will rest on no general analysis of that notion.

I *have* got an analysis of determinism. But I have given this analysis (in various more or less equivalent forms) elsewhere and I shall not repeat it here.[3] I will remark, however, that determinism is the thesis that the past and the laws of nature together determine a unique future and is *not* the thesis that every event has a cause ("universal causation"). For the thesis of universal causation might be true and determinism false.[4]

3. See the papers referred to in note 1.
4. See pp. 89 and 90 of my "Reply to Narveson," *Philosophical Studies* 32 (1977).

However the thesis of determinism (the thesis that the past and the laws of nature determine a unique future) should be spelled out in detail, it should have the following consequence. (In the sequel I shall, in order to save space, conflate use and mention to a really *shocking* extent. You have my word for it that this conflation is eliminable by dull and lengthy paraphrasis.) Let S be a sentence that, in some relevant sense, gives a complete and accurate description of the entire state of the world at some moment in the remost past. In fact, let us suppose that S gives a description of the state of the world at some moment so long ago that at that moment there were no human beings and never had been any. (It will facilitate the argument to suppose there was such a moment. This assumption could be dispensed with at the cost of uninteresting complications.) Let L be a sentence that, in some relevant sense, gives a complete and accurate statement of "the laws of nature," whatever, precisely, those may be. Let T be any truth whatever. Let '$\Box$' represent what Plantinga has called "broadly logical necessity," that is, truth in all possible worlds. Then it follows from determinism that

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

It is this consequence of determinism that I shall show is incompatible with moral responsibility.

II

I shall use 'Np' as an abbreviation for the following sentence form:

> p and no human being, or group of human beings, is even partly responsible for the fact that p.

For example, 'N Nixon received a pardon' is to be read, 'Nixon received a pardon and no human being or group of human beings is even partly responsible for the fact that Nixon received a pardon.' The qualification introduced by the words 'even partly' will play no role in the argument of this paper and I shall ignore it in the text. The curious reader may consult footnote 5. Owing to the presence of the word 'human' in this sentence-form, my arguments will be directly applicable only to questions of human moral responsibility. I have included the word 'human' in order to avoid discussing the relation between determinism and the actions of supernatural agents such as

God or angels. The argument of the sequel, however, could easily be applied to Martians, Venerians, or any other purely natural agents.

My argument will make use of two inference forms involving 'N':

(A) $\Box p \vdash \mathrm{N}\, p$

and

(B) $\mathrm{N}p, \mathrm{N}(p \supset q) \vdash \mathrm{N}q$.

The validity of (A) seems to me to be beyond dispute. No one is responsible for the fact that $49 \times 18 = 882$, for the fact that arithmetic is essentially incomplete, or, if Kripke is right about necessary truth, for the fact that the atomic number of gold is 79. (According to Descartes, God is responsible for these things; but we needn't consider that vexed question.) The validity of (B) is a more difficult matter. I shall return to it later.[5]

My argument will require two premises, 'NS' and 'NL'. The former is obviously true, since no human being is morally responsible for anything that occurred before any human beings had ever been. The latter is obviously true, since, whatever may be true of God or other supernatural beings, no human being is morally responsible for the laws of nature. (For example, if it is a law of nature that nothing travels faster than light, then no human being is morally responsible for the fact that nothing travels faster than light.)

Now the argument. We begin with our consequence of determinism:

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

From (1) we may deduce by elementary modal and sentential logic,

5. If the words 'even partly' were omitted from the sentence-form that 'N*p*' abbreviates, then (B) might be open to counterexample. Suppose, for example, that Smith kills the elder of the Jones twins and that the younger is killed by a bolt from the blue. It is at least arguable that in that case neither Smith nor anyone else is responsible for the fact that *both* the Jones twins are dead. But then the following argument has true premises and a false conclusion

N Both the Jones twins are dead
N (Both the Jones twins are dead $\supset$ the elder of the Jones twins is dead)
$\therefore$ N The elder of the Jones twins is dead

if the words 'even partly' are omitted from the reading of 'N*p*'. But it seems evident that, in the case imagined, Smith is at least *partly* responsible for the fact that both the Jones twins are dead.

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

We now argue:

(3) $N(S \supset (L \supset T))$	From (2) by (A)
(4) NS	Premise
(5) $N(L \supset T)$	From (3) and (4) by (B)
(6) NL	Premise
(7) NT	From (5) and (6) by (B)

I have called this an argument. More precisely it is an argument-form. We may derive indefinitely many arguments from it by substituting arbitrary sentences for 'T'. If we substitute for T a sentence that expresses a truth and if determinism is true, the substitution-instance of (1) so obtained will be true and the argument so obtained will be sound (assuming, of course, that it is valid). This fact about our argument-form amounts to a proof of the following proposition: substitute any *truth* you like for 'T' in the following schema

> If determinism is true, then no human being, or group of human beings, is morally responsible for the fact that T,

and you will get a truth. For example, if you substitute 'Kennedy was assassinated', 'The U.S. used atomic weapons against Japan', or 'Nixon received a pardon' for 'T', you will get a truth. This result, I think, may be properly summarized in these words: determinism is incompatible with moral responsibility.

We have proved this result provided that the reasoning employed in our argument-form is valid; that is, provided that both (A) and (B) are valid; that is—since the validity of (A) is beyond dispute—provided (B) is valid. Let us now turn to the question of the validity of (B).

III

How could one show that (B) is valid? How, in general, does one go about showing that an argument-form is valid? There would seem to be two ways.

First, one might employ the methods of formal semantics. In the present case, since 'N' is very like a modal operator, the methods of *possible-world* semantics might seem promising. Here is a sketch of how we might apply these methods to (B). We first delimit a certain set W

of worlds and say that N*p* is true just in the case that *p* is true in all these worlds. (This would amount to a semantical definition of 'N'.) For example, we might say that N*p* is true if *p* is true in both the actual world and in all worlds such that human beings can be held morally responsible for their "actuality-status" (that is, actuality or non-actuality, as applicable). Interestingly enough, the definition of *W* is of no *formal* significance. If we accept any definition of N*p* of the following form: 'Np is true iff p is true in all worlds such that . . . ', where the condition that fills the blank makes no mention of *p*, then (B) will "come out" valid. (Obviously, if p is true in every member of W, and if $p \supset q$ is true in every member of *W*, then *q* is true in every member of *W*.) While this formal result is not devoid of persuasive force (despite its utter triviality), it is far from decisive. It depends on the assumption that there is *some* set of worlds *W* such that N*p* *can* plausibly be thought of as making the assertion that *p* is true in every member of *W*. While this assumption seems right to me, I have no argument for it, and a person who was determined to reject (B) might very well reject it.

Secondly, one might attempt to show that (B) was valid by "reducing" it to certain generally accepted valid inference-forms. But it seems intuitively evident that this cannot be done. No generally accepted inference-form involves moral concepts. (The familiar principle that 'ought' implies 'can' may be an exception to this generalization. But even if this principle does count as a "generally accepted inference-form," it's hard to see how it could be of much help to the friends of (B).) And it seems wholly implausible to suppose that an inference-form essentially involving the concept of moral responsibility could be reduced to inference-forms involving only non-moral concepts.

Thus the prospect of *showing* (B) to be valid appears bleak, though perhaps no bleaker than the prospect of *showing* anything of philosophical interest. I must confess that my belief in the validity of (B) has only two sources, one incommunicable and the other inconclusive. The former source is simply what philosophers are pleased to call "intuition": when I carefully consider (B), it seems to be valid. But I can't expect *you* to be very impressed by this fact. People's intuitions, after all, have led them to accept all sorts of crazy propositions, and many sane but false propositions. (The Unrestricted Comprehension Principle in set theory and the Galilean Law of the Addition of Velocities in physics are good examples of propositions in the second category.) The latter source is the fact that I can think of no instances of (B) that have, or could possibly have, true premises and a false

conclusion. That is, I can think of no instances of (B) that can be seen to have true premises and a false conclusion *independently* of the question whether moral responsibility is compatible with determinism. If moral responsibility is compatible with determinism (and if determinism is true), then the following instance of (B):

N(S & L. ⊃ The U.S. used atomic weapons against Japan)
N(S & L)
∴ N The U.S. used atomic weapons against Japan

doubtless has true premises and a false conclusion.

It may be hard to credit, but there are almost certainly philosophers who would say that this shows that my use of (B) "begs the question" against the proponents of the compatibility of determinism and moral responsibility. But if this accusation of question-begging were right, it's hard to see how any argument could avoid begging the question. If one presents an argument for a proposition Q, then, if Q is false, *some* step in the argument is wrong; and one may believe of a certain step in the argument that *if* any step is wrong, *that one* is. But it hardly follows that one is "begging the question" by taking that step. One may be begging the question (whatever, precisely, that is) but that one is begging the question is not a consequence of the mere existence of a "weakest link" in one's chain of reasoning.

But these questions about "question-begging" and where the burden of proof lies, and so on, are very tricky. Let's look at them from a different angle. Suppose a proponent of the compatibility of determinism and responsibility (let's call this doctrine R-compatibilism) replies to my argument as follows: "You employ argument-form (B). But this argument-form is invalid. I prove this as follows:

R-compatibilism is true
∴ Argument-form (B) is invalid.

You yourself admit that the conclusion of this argument follows from its premise [I do]. You may not accept its premise, but that's *your* problem, for that premise is *true*. Moreover, you can hardly object to this little argument of mine on the ground that it begs the question. It's no worse in that respect than *your* argument, which is essentially this:

Argument-form (B) is valid
∴ R-compatibilism is false."

What am I to say to this? I suppose I can do no more than appeal to the intuitions of my audience. Here's how it looks to me (and doesn't it look this way to you?): Argument-form (B) seems obviously right and R-compatibilism does not seem obviously right. If two principles are in conflict and one seems obviously right and the other does not seem obviously right, then (if one must choose) one should accept the one that seems obviously right.

But perhaps someone will say that he finds R-compatibilism obviously right. Presumably this attitude of his is either grounded in an immediate and intuitive relationship to R-compatibilism—he claims to *see* that it's true, just as I claim to see that (B) is valid—*or* his attitude is grounded in some argument for R-compatibilism. Let us first look at the case of the philosopher who claims to see the truth of R-compatibilism intuitively. Well, arguments, like explanations, must come to an end somewhere. Perhaps if there is such a philosopher, he and I constitute a genuine case of a conflict of rock-bottom intuitions. But I must say I should find any such claim as the one I have imagined incredible. R-compatibilism looks to me like the kind of thing one could believe only because one had an argument for it. I simply cannot see what could be going on in the mind of someone who claimed to know it intuitively. I don't know what that would *feel* like.

The philosopher who believes R-compatibilism on the basis of an *argument* is not likewise mysterious to me. But I shall want to know what the premises of his argument are. And I shall raise the following question about his (ultimate) premises: Are they *really* intuitively more plausible than (B)? I find it hard to believe that there are any propositions that entail R-compatibilism that are more plausible than (B). I'm not sure what premises might be employed in an argument for the compatibility of responsibility and determinism, but I know what the premises employed in arguments for the compatibility of *free will* and determinism are like and I expect that the premises of arguments for R-compatibilism would be of a comparable level of plausibility. The crucial premise in arguments for the compatibility of free will and determinism is usually a semantic proposition that begins in some such way as this

'S can do A' means 'S would do A if S chose to do A and . . .'

and ends in complexity.[6] When I examine premises of this sort, I find

6. See, e.g., Wilfrid Sellars, "Fatalism and Determinism," in Keith Lehrer, ed., *Freedom and Determinism* (New York: Random House, 1966); Bernard Gert and Timothy J. Duggan, "Free Will as the Ability to Will," Chapter 10 in the present volume; Keith Lehrer, "Preferences, Conditionals and Freedom," in Peter van Inwagen, ed., *Time and Cause: Essays Presented to Richard Taylor* (Dordrecht: D. Reidel, 1980).

myself without any particular convictions about their truth or falsity, owing simply to their complexity. If someone presents an argument for R-compatibilism that has a premise as complex as any of these semantical premises that figure in the free-will debate, then naturally I shall find this complex premise less plausible than (B) and will continue to accept (B) and its consequences, among which is R-incompatibilism.

No one, of course, is obliged to correct my mistaken beliefs. But if anyone thinks my belief in R-incompatibilism is false and *does* for some reason take an interest in my intellectual welfare, here is what he will have to do to get me to see the light: he will have to produce some proposition intuitively more plausible than the proposition that (B) is valid and show that this proposition entails R-compatibilism, or else he will have to devise a counterexample to (B) whose status as such can be established without assuming that determinism and moral responsibility are compatible.