The Philosophical Quarterly, Vol. 47, No. 188 July 1997
ISSN 0031–8094

CRITICAL STUDY

FISCHER ON MORAL RESPONSIBILITY

By Peter van Inwagen

The Metaphysics of Free Will: an Essay on Control. By John Martin Fischer. (Oxford: Blackwell, 1994. Pp. ix + 273. Price not given.)

That moral responsibility entails indeterminism is not an attractive thesis. Anyone who accepts this thesis must be willing to concede that, since determinism could turn out to be true, our deeply ingrained conviction of the reality of moral responsibility could turn out to be an illusion. But this unattractive thesis is a logical consequence of two very plausible propositions:

> Free will (that is, the ability to act otherwise than one in fact does) cannot exist in a fully deterministic world
>
> Moral responsibility requires free will: if one cannot ever act otherwise than one does, then one is morally responsible for none of the consequences of one's acts.

Plausible as these propositions are, neither is so evident that it cannot be denied. If, like Hobbes, Hume and Mill, one denies the first, one embraces *compatibilism*. But compatibilism is nowadays widely regarded as implausible, owing to the fact that compatibilists must deny a very plausible thesis that I shall call the principle of the transfer of inability (PTI). One way of formulating PTI is as the thesis that the following rule of inference is valid:

> It is true that p, and A is unable to bring about the falsity of this proposition
>
> If it is true that p, then it is true that q, and A is unable to bring about the falsity of this (conditional) proposition

hence

> It is true that q, and A is unable to bring about the falsity of this proposition.

And it does seem very plausible indeed to suppose that this rule is valid. (The following informal argument shows that the validity of PTI entails incompatibilism. Let p_0 be a proposition expressing the state of the world at some moment in the remote past, and let p be a proposition expressing the present state of the world. Then, if determinism is true, p_0 and the laws of nature together entail p. But entailments are necessary truths, and no one is able to bring about the falsity of a necessary truth. Furthermore, no one is able to bring about the falsity of either p_0 or any law of nature. It follows, by PTI, that no one is able to bring about the falsity of p. This informal argument can easily be formalized, and the validity of the resulting formal argument can easily be seen to depend only on the principles of standard logic, PTI, and the principle that from the premise that a given proposition is a necessary truth, the conclusion follows that no one is able to bring about its falsity.)

If one is not a compatibilist – either because one accepts the principle of the transfer of inability or for some other reason – must one then concede that moral responsibility cannot exist in a fully deterministic world? This may be said to be the central question of John Martin Fischer's *The Metaphysics of Free Will*. (But this statement needs to be qualified. The book is only partly devoted to questions about what could be true in a deterministic world. It is also partly devoted to questions about what could be true in a world in which God had perfect knowledge of the future actions of human beings. I shall not discuss this aspect of the book.) Fischer's answer is 'No', for he holds that the second of our 'two very plausible propositions' is false: moral responsibility does not require free will. Although he defends a wide variety of theses in *The Metaphysics of Free Will* (e.g., that the principle of the transfer of inability does not entail, as I have argued it does, that the ability to do otherwise is rare; that the solution to Newcomb's Problem depends on whether the predictor is 'infallible' or merely 'inerrant'), the following three theses are, in my judgement, the core theses of the book:

> It is at least very likely that free will is incompatible with determinism (and, therefore, those who believe in moral responsibility would be ill advised to allow their case to rest on compatibilism)
>
> Examples of the kind devised by Harry Frankfurt in his classic essay 'Alternate Possibilities and Moral Responsibility' show that moral responsibility does not require free will (that morally responsible agents may be without the power to act otherwise than they do)
>
> Although moral responsibility does not require free will, it does require a certain sort of control over one's actions; but the sort of control it does require is compatible with determinism.

I shall make some brief remarks about the first thesis, and then go on to discuss the second at some length. I shall, finally, offer a short criticism of the third thesis.

Although Fischer thinks that there are very plausible arguments for the conclusion that free will is incompatible with determinism, he holds that arguments for this conclusion need not appeal to the principle of the transfer of inability (or the principle of the transfer of powerlessness, as he calls it), and that in fact the *most*

plausible argument for the incompatibility of free will and determinism does not appeal to PTI. The most plausible argument is this: as Carl Ginet has said (and this is very well said indeed) 'freedom is the freedom to add to the actual past'; and any 'addition to the actual past' that anyone – anyone who is not a *bona fide* miracle-worker – is able to make must be causally continuous with the actual past; but if the world is fully deterministic, the only possible additions to the actual past that are causally continuous with the actual past are the additions that are actually made. That this powerful little argument does not depend on PTI (and is more plausible than any argument for the incompatibility of free will and determinism that does depend on PTI) is a very interesting contention, and it is important if it is true. It is, moreover, only one of a great many closely related conclusions about determinism, free will and PTI that Fischer attempts to establish in (roughly) the first half of the book. But, important and interesting as these conclusions are, his conclusions about the relation between free will and moral responsibility are even more important and interesting, and I shall devote the body of this discussion to them.

Fischer's arguments for these conclusions are challenging, and anyone who is interested in the relation between moral responsibility and the ability to do otherwise will have to take account of them. They are, in my judgement, the most important arguments of the book, the arguments on the basis of which, in the last analysis, the importance of the book's contribution to our understanding of the problem of free will and moral responsibility must be evaluated.

Perhaps it is unsurprising that I have not been convinced by these arguments, for they go contrary to some long-standing convictions that I brought to my reading of the book. I shall try to explain why I have not been convinced and the reader may judge. In my view, the conceptual issues raised by the 'Frankfurt-style examples' on which Fischer's arguments turn are of extreme delicacy, and the language that Fischer employs in his discussion of them is insufficiently precise to do justice to this delicacy. The remainder of this study is largely an elaboration of this contention.

Everyone, I think, will agree that examples of the sort that Frankfurt employed in 'Alternate Possibilities and Moral Responsibility' (those remarkable examples involving the potential but not actual manipulation of an agent) are of the first importance for an understanding of the relation between free will and moral responsibility. But how, exactly, are they to be used? How should they be deployed in argument? What is their *point*? Fischer generally talks as if reflection on Frankfurt-style examples can be used to establish some positive conclusion about responsibility and the ability to do otherwise. For example (p. 158):

> [Frankfurt-style examples] point us to something both remarkably pedestrian and extraordinarily important: moral responsibility for action depends on what actually happens. That is to say, moral responsibility for actions depends on the actual history of an action and not upon the existence or nature of alternative scenarios.

This strikes me as at best misleading. There are various principles that, given the premise that we are unable to do otherwise, enable us to deduce the conclusion that we lack moral responsibility. The question should be: are Frankfurt-style examples *counter-examples* to these principles? One could of course say in Fischer's defence that

the passage I have quoted (and the same could be said of many similar passages) implies that Frankfurt-style examples have just this property. In the quoted passage, Fischer clearly means to imply that Frankfurt-style examples, or some of them, are counter-examples to some such principle as 'Moral responsibility for an action depends not only on the actual history of that action, but also on the existence of alternative scenarios of a certain nature'. I concede that this passage does have this implication, but the principle I have extracted from it is, in my view, too vague for a useful discussion to be possible of the question whether it is refuted by Frankfurt-style examples. There are, moreover, relatively precise principles relating moral responsibility and the ability to do otherwise that are *not* refuted by Frankfurt-style examples – or so I have argued, and nothing Fischer has said in this book has led me to second thoughts about my arguments. Here (I contend) is such a principle:

> If it is a fact that *p*, an agent is morally responsible for the fact that *p* only if that agent was once able to act in such a way that it would not have been the case that *p*.

It is important to remember that, however many *other* principles relating free will and moral responsibility there may be that can be shown to be false by Frankfurt-style examples, if *this* principle is true, then no agent who is unable ever to act otherwise is morally responsible for any fact. And if no agent is morally responsible for any fact, then, it would seem, our belief that there is such a thing as moral responsibility is illusory. The same point can be made about any other principle that implies that moral responsibility requires free will. In the end, Frankfurt-style examples will be of little interest unless they can be used to refute *all* principles that imply that moral responsibility requires free will.

Can Frankfurt-style examples be used to show that my 'relatively precise' principle is false? Let us try to construct one.

> Cosser wanted Gunnar to shoot and kill Ridley, which Gunnar seemed likely to do; he intended to, and he had the means and the opportunity. But if Gunnar had changed his mind about killing Ridley, Cosser would have manipulated Gunnar's brain in such a way as to have re-established his intention to shoot Ridley. In the event, Cosser's 'insurance policy' turned out not to have been necessary, for Gunnar did not change his mind, and shot and killed Ridley 'on schedule'. Cosser played no causal role whatever in the sequence of events that led up to the killing.

Have we a counter-example to our principle? Before we can say that we have, we must find some appropriate sentence to replace '*p*' in the principle. Let us suppose that, Ridley having been a widower, his children are now orphans. There was, moreover, no 'second gunman', and no there was no fatal heart attack or car crash lurking nearby in logical space: if Gunnar had not shot Ridley, Ridley's children would *not* now be orphans. Having made these stipulations, let us replace '*p*' with 'Ridley's children are now orphans'.

Was Gunnar able to act in such a way that, if he had, Ridley's children would not now be orphans? It would seem not, for if he had changed his mind and decided

not to shoot Ridley (assuming that he was able to change his mind), Cosser would have 'changed his mind back', and he would have killed Ridley anyway: in every future that was open to Gunnar from the moment Cosser established his 'insurance policy', Gunnar killed Ridley. (Let us ignore the fact that our story leaves open the possibility that there was some earlier moment at which a future in which he did not kill Ridley was open to Gunnar.)

Is Gunnar morally responsible for the fact that Ridley's children are orphans? 'Of course he is', Frankfurt and his followers argue. 'Look, suppose you subtracted Cosser from the story. Let us call the story of Gunnar and Ridley *sans* Cosser "the truncated story". In the truncated story, Gunnar is obviously morally responsible for the fact that Ridley's children are orphans – at least if moral responsibility is possible at all. (If you think that something special has to be added to the truncated story to ensure that Gunnar is responsible for this fact – indeterminism, "agent causation" – feel free to add it.) Now suppose Cosser is put back into the story. Does Cosser's re-entry into the story absolve Gunnar of the responsibility that was his in the truncated story? How could it? In the story in which Cosser once again figures, Cosser was waiting in the wings all the while, but he *did* nothing, or nothing that affected Gunnar; everything in, say, a mile-wide region of space-time centred on Gunnar's space-time trajectory (up to the moment he pulled the trigger) was just as it would have been if Cosser had never existed. And surely, if Gunnar's pulling the trigger made it causally inevitable that Ridley's children are now orphans, ought we not to be able to settle the question whether Gunnar was morally responsible for the fact that those children are now orphans by examining nothing but the content of this region?'

Many find this style of reasoning incontrovertible. It must be remarked, however, that the state of things outside a region of space-time can have important consequences for what is true of things inside that region. After all, adding Cosser and his powers and his dispositions to employ them to the truncated story changes the truth-value of

> Gunnar was able to act in such a way that, if he had, Ridley's children would not now be orphans

from true to false. Why cannot adding Cosser to the truncated story do the same for

> Gunnar is morally responsible for the fact that Ridley's children are now orphans?

The suggestion that the addition of Cosser has this consequence is likely to be met with incredulous stares. But why would it not be appropriate to confront the corresponding suggestion about Gunnar's abilities with the same stares? How, one might ask (staring incredulously), could something that in no way affects one's body, mind or immediate environment – that in no way affects the content of the region of space-time that surrounds one – have any effect on one's *abilities*?

It might be worth-while to take this question seriously and to try to answer it. The answer is: well, in a way it cannot – it cannot diminish one's skill as a marksman, or make one any less a master of disguise, or diminish one's physical courage

or one's reaction time; but it *can*, as we have seen, affect one's abilities with respect to determining the truth-values of various propositions.

Similarly, I would say, factors that have no effect on an agent's body, mind or immediate environment can be among the factors that determine whether the agent is morally responsible for certain facts. If it would have been the case that *p* no matter what choices or decisions Alice had made (provided only that she made them 'on her own', without having been caused to do so by some 'outside' agency), then it seems plausible to suppose that Alice could not be morally responsible for the fact that *p*. This principle – let us call it the 'no matter what' principle – is extremely attractive, and, to my mind, Frankfurt-style examples do nothing to lessen its attractiveness. That Ridley's children are orphans is a fact. If Ridley's children would have been orphans if Gunnar had decided 'on his own' not to shoot Ridley – if they would have been orphans no matter what he had decided on his own – then how can he be morally responsible for the fact that they are orphans?

Or do Frankfurt-style examples simply show that the 'no matter what' principle is false? If they do, then, I think, it could be shown to be false by much simpler cases than those Frankfurt has constructed (simpler because they do not involve off-stage potential manipulators). But, I would argue, these simpler cases do not refute the 'no matter what' principle, and, when one compares these simple cases with 'potential manipulator' cases, one will note that the potential manipulator adds nothing of philosophical relevance to what is contained in the simple cases. Here is one:

> I am supposed to take the serum upriver to the plague-stricken village. But I get drunk and miss the boat. Taking the boat is the only possible way to get to the village. Soon after the boat leaves the dock, it strikes a rock and sinks. Hundreds of villagers who would have been saved by the serum die.

Here is a fact: hundreds of villagers do not get the serum and consequently die. Am I morally responsible for this fact? My own reaction to this question is simple and unequivocal: of course not. And the reason is that the villagers would have died no matter what choices or decisions I had made; in particular, if I had chosen to remain sober, and had made every possible effort to ensure that the serum reached the village, the villagers would still have died. If I am charged with the deaths of the villagers, I have a perfect excuse: it was not possible for me to save them. Of course, no one is likely even to consider holding me morally responsible for *that* fact. If the story comes out, my superiors will hold me guilty of dereliction of duty, and I shall no doubt not be trusted with anything of any importance again; I shall no doubt be a moral pariah. I shall very likely be told that I behaved 'irresponsibly'. All this is without doubt. But these things that cannot be doubted do not change the fact that I am not responsible for the deaths of the villagers. It is true that if I tried to defend myself by saying something along the lines of 'But they would have died even if I had stayed sober and been on the boat, so I'm not responsible for their deaths', this will be universally received as a contemptible attempt to defend the indefensible. But all that that shows is that making a true statement can, in certain circumstances, be a contemptible attempt to defend the indefensible. (And this we already knew: those who say 'I didn't mean to' are usually speaking the truth.)

I do not see why we should not respond to Frankfurt cases proper (potential-manipulator cases) in the same way. For what it is worth, and it is not worth much, Gunnar is not morally responsible for the fact that Ridley's children are orphans. (There are, of course, lots of facts he *is* morally responsible for – that he shot Ridley without having been caused to do so by Cosser, for example, or that he did not even try to avoid shooting him.) It does not follow, however, that it is improper for Ridley's children to hold him responsible for the events of that terrible day on which they became orphans, for there is more to moral responsibility than responsibility for facts (or than moral responsibility for the truth-values of propositions).

The analogy of the legal determination of guilt and innocence is instructive. Here is a chestnut. Jane plans to go for a long trek in the desert. Poisson and Sandy both desire her death. Poisson poisons her water-bottle. Sandy, not knowing what Poisson has done, empties the water-bottle and fills it with sand. As a result, Jane dies of thirst in the desert. The facts come out, Poisson and Sandy are arrested, and Poisson is convicted of attempted murder and Sandy of murder. (Sandy's defence, that he in fact *extended* Jane's life by removing the poisoned water from her water-bottle and replacing it with harmless, if useless, sand, is laughed out of court.) Why? Because Sandy caused Jane's death; he caused *the death that Jane in fact died*. The question the court considers is not 'Who caused the proposition that Jane died in the desert on or about 12 July to be true?'. The question is rather 'Who caused Jane's death?', a question about a concrete, individual event. But this does not mean that all questions about the causation of facts or of the truth-values of propositions are irrelevant to the court's deliberations, for it is obvious that in causing any event one must cause certain facts to obtain. (For example, Sandy could hardly have caused 'the death that Jane in fact died' if he had not caused it to be the case that her water-bottle was filled with sand.)

The points I have made are about causation rather than responsibility, but causation and responsibility are not unconnected notions. It seems to me to be evident that Sandy did not cause it to be the case that Jane died in the desert on or about 12 July, for Jane would have died in the desert on or about 12 July no matter what choices or decisions Sandy had made. And it seems to me to be evident, for exactly the same reason, that Sandy was not morally responsible for Jane's having died in the desert on or about 12 July. But he *was* morally responsible for her death (or he was if anyone is morally responsible for anything). And he could not have been morally responsible for her death if he had not been morally responsible for some of the facts relating to her death – such as the fact that her water-bottle contained only sand. If, therefore, one decides on general philosophical grounds that Sandy was unable to act otherwise than he did – courts are not philosophical seminars; courts simply take it for granted that people are in general able to act otherwise, just as they simply take it for granted that sense-perception is in general reliable – then one should conclude that he was morally responsible for no facts relating to the case; and one should go on to conclude that because he was morally responsible for no facts relating to the case, he was therefore not morally responsible for Jane's death (or for any other concrete event or for anything whatever).

I have tried to show why I remain unconvinced by Fischer's attempt to show that moral responsibility does not require free will. My explanation has consisted entirely of very well known considerations, but, in my view, nothing Fischer says renders these old considerations any less effective. Indeed, his arguments are not clearly addressed to these considerations. Fischer's arguments are addressed to very 'broad' questions that he formulates by means of abstract nouns (for example, 'What is the relation between moral responsibility and alternative possibilities?'). As I see the problem of the relation of moral responsibility to free will, this problem is so subtle and complex that a useful discussion of it must take the form of an attempt to answer some very narrow questions about precisely formulated principles.

I wish, finally, to make a brief point about the kind of 'control' that, according to Fischer, *is* necessary for moral responsibility. Fischer holds, moreover, that this sort of control is the *only* sort of control that is necessary for moral responsibility. In fact, if I understand him, he maintains that exercising this sort of control over one's actions is not only necessary but *sufficient* for being morally responsible for them. (I have always deprecated talk of being morally responsible for one's actions. In my view, we hold people morally responsible for the results or consequences of their actions, not for the actions themselves. But I do not insist on this point here.) Here is a somewhat condensed statement of Fischer's position: an agent is morally responsible for his actions if and only if those actions issue from internal decision-making mechanisms that are 'weakly responsive to reasons'. That is, an agent who performs some act is morally responsible for that act if and only if, if the agent's internal decision-making mechanisms (which in actuality issued in a decision *to* perform the act) had been just as they in fact are, and if they had received as 'input' some realization or discovery that, in the circumstances, would constitute what they would interpret as a good reason not to perform that act, their operations would have resulted in the agent's deciding *not* to perform that act.

If this thesis about moral responsibility is correct, then it is obvious that moral responsibility is compatible with determinism. (And if the ability to do otherwise is incompatible with determinism, then moral responsibility is compatible with an inability to do otherwise.) But is it correct? It would seem not. Suppose a paranoid schizophrenic murdered a stranger, believing that the stranger was an agent of the evil king of Pluto – a paradigm case, surely, of someone who is not morally responsible for what he has done. But the internal decision-making mechanisms of this madman were no doubt weakly responsive to reasons: if someone had stepped up to him just as he was drawing his knife and had whispered, 'Jorkins, MI5. Don't kill him. We're tailing him to find out who he reports to. We have a more important mission for you. Go to this address and knock three times', he would no doubt have decided not to murder the stranger. He is therefore, if Fischer is right, morally responsible for having killed the stranger. Something has obviously gone wrong. Curiously enough, Fischer is aware of examples like this one (see p. 243 fn. 8), but says only that, although such cases do show that his thesis needs to be revised, he is hopeful that the required revision will not be radical and that it will leave his essential point intact. I think that many readers will share my reaction to this statement: we shall want to see the revision before we agree that moral responsibility

requires no more 'control over one's actions' than is provided by some sort of potential responsiveness to reasons (and we shall insist that Fischer really ought to have dealt with cases like the 'madman' at length in the text and not simply by issuing a brief promissory note in small print at the back of the book).

I have tried to explain why I have not been convinced by Fischer's arguments. But it was hardly to be expected that I should have been convinced by them, for Fischer's conclusions are inconsistent with theses I have defended for many years. More open-minded readers may be convinced by Fischer's carefully stated and well organized arguments. *The Metaphysics of Free Will*, whether its conclusions are right or wrong, is an important contribution to the problem of free will and moral responsibility.

The University of Notre Dame