# DISCUSSION

## VAN INWAGEN ON FREE WILL

### BY JOHN MARTIN FISCHER

Recently, Peter Van Inwagen has presented a family of three arguments for the incompatibility of causal determinism and human freedom.[1] Here, I wish to focus on what he calls the "First Formal Argument". I shall contend that it is not convincing because it employs a premise which is highly controversial, and the evidence adduced in support of this premise does *not* support it over weaker premises which are consistent with compatibilism. I do not plan to argue that Van Inwagen's argument is obviously unsound; rather, I intend to show that it is not decisive, because it is incomplete in an important sense.

## I

### VAN INWAGEN'S ARGUMENT

The argument can be presented as follows. Determinism is defined as the conjunction of the following two theses:

> For every instant of time, there is a proposition that expresses the state of the world at that instant;

> If $p$ and $q$ are any propositions that express the state of the world at some instants, then the conjunction of $p$ with the laws of nature entails $q$. (p. 65)

Also, the following symbols are used. '$T_0$' shall denote some arbitrarily chosen instant of time earlier than J's birth; '$P_0$' shall denote the proposition that expresses the state of the world at $T_0$; 'P' shall denote the proposition that expresses the state of the world at T (it implies that J did not raise his hand at T), and 'L' shall denote the conjunction into a single proposition of all the laws of nature (pp. 69–70).

Van Inwagen's "First Formal Argument" consists of seven propositions, the seventh of which is alleged to follow from the first six (p. 70):

> (1) If determinism is true, then the conjunction of $P_0$ and L entails P.
> (2) It is not possible that J have raised his hand at T and P be true.
> (3) If (2) is true, then if J could have raised his hand at T, J could have rendered P false.

---

[1] Peter Van Inwagen, *An Essay on Free Will* (Oxford, 1984); page references hereafter are to this work. Van Inwagen first presented similar arguments in: "A Formal Approach to the Free Will Problem", *Theoria* 40 (1974), and "The Incompatibility of Free Will and Determinism", *Philosophical Studies* 27 (1975), pp. 185–99.

(4) If J could have rendered P false, and if the conjunction of P₀ and L entails P, then J could have rendered the conjunction $P_0$ and L false.

(5) If J could have rendered the conjunction of $P_0$ and L false, then J could have rendered L false.

(6) J could not have rendered L false.

So (7) If determinism is true, J could not have raised his hand at T.

This argument is powerful and disturbing. In evaluating it, it is crucial to understand the key phrase 'S can render P false'. Consider the following account: "S can render P false just in case it is within S's power to arrange or modify the concrete objects that constitute his environment in some way such that it is not possible "in the broadly logical sense" that he arrange or modify these objects in that way and P be true" (p. 67). Roughly speaking, I can render P false, on this account, insofar as I can affect my environment in such a way that my doing so would *entail* that P is false. Let us call this definition "account I".

Although Van Inwagen earlier presented this as adequately capturing the intuitive concept of being able to render a proposition false, he now recognizes that this account is unacceptable.[2] He says:

> Let us suppose that in 1550 Nostradamus predicted that the Sphinx would endure till the end of the world. And let us suppose that this prediction was correct and, in fact, that *all* Nostradamus's predictions were correct. Let us also suppose that it was within Gamel Abdel Nasser's power to have the Sphinx destroyed. Then, I should think, it was within Nasser's power to render false the proposition that all Nostradamus's predictions were correct. But this would not be the case according to the definition proposed in the preceding paragraph [account I], since it is possible in the broadly logical sense that Nasser have had the Sphinx destroyed and yet all Nostradamus's predictions have been correct. That is, there are possible worlds in which the proposition that all Nostradamus's predictions were correct is true and in which Nasser had the Sphinx destroyed: worlds in which Nostradamus did not predict that the Sphinx would endure till the end of the world and made no other predictions that would have been falsified by Nasser's destruction of the Sphinx. (pp. 67–68)[3]

In order to rule out such counter-examples, Van Inwagen suggests that we "build the past into" our definition, suggesting that we define 'S can render P false' as follows:

> It is within S's power to arrange or modify the concrete objects that constitute his environment in some way such that it is not possible in the broadly logical sense that he arrange or modify these objects in that way and the past have been exactly as it in fact was and P be true. (p. 68)

Let us call this definition "account II". Van Inwagen does not claim that this is the *only* way of ruling out the counter-examples; rather, he says that it seems to be the

[2] Account I is the proposal which Van Inwagen accepted in his "Reply to Narveson", *Philosophical Studies* 32 (1977), p. 93.

[3] For a similar criticism of a parallel suggestion made by Joshua Hoffman, see: John Martin Fischer, "Freedom and Foreknowledge", *Philosophical Review* 92 (1983), p. 72.

"best way" to do so. It should be noted that this account of 'S can render P false' has the consequence of making premise (5) trivially true – (5) follows directly from account II of 'S can render P false', and Van Inwagen exploits this fact in his presentation of his argument.

## II

### THE DENIAL OF PREMISE (6)

Whereas premise (5) is trivially true, on account II, premise (6) is controversial. Consider what Van Inwagen says about (6): "This premise would seem to be an obvious consequence of what we said about our powers with respect to the laws of nature . . . " (p. 72). His claim about the relationship between our powers and the laws of nature can be understood on the basis of the following sort of example:

> Suppose a bureaucrat of the future orders an engineer to build a spaceship capable of travelling faster than light. The engineer tells the bureaucrat that it's a law of nature that nothing travels faster than light. The bureaucrat concedes this difficulty, but counsels perseverance: 'I'm sure', he says, 'that if you work hard and are very clever, you'll find some way to go faster than light, even though it's a law of nature that nothing does'. Clearly his demand is simply incoherent. (p. 62)

Van Inwagen's point is that if performing a certain act would require that a law of nature be violated, then no human agent can perform the act. But his defence of the controversial premise (6) is a bit misleading, and, I believe, unsatisfying. It is misleading because one might deny (6) for reasons *entirely apart* from any consideration of the relationship between our powers and natural laws; that is, one might *accept* everything Van Inwagen says about the relationship between natural laws and human freedom and still deny (6).

Let us suppose that one believes that in the situation described in Van Inwagen's argument, the following "backtracking conditional" would be true: If J had raised his hand at T, then $P_o$ wouldn't have obtained at $T_o$. (I shall consider the position which *denies* this backtracker below.) The conditional here is to be interpreted *non-causally*. Now, if one also believes that J could at T have raised his hand at T, then one is committed to the following conjunction of a "can-claim" and a backtracking conditional: "J could at T have raised his hand at T, and if he had done so, the past would have been different from what it actually was (insofar as $P_o$ wouldn't have obtained at $T_o$)". It follows that one believes that J could at T have rendered L false (on account II), because one believes that J could have so arranged things (raised his hand) that it was impossible that he so arrange things and the past have been as it actually was (i.e., $P_o$ obtained at $T_o$) and L be true. But since one believes that if J had raised his hand at T, then $P_o$ wouldn't have obtained at $T_o$, one *doesn't* believe that if J had raised his hand at $T_o$, L wouldn't have obtained.

So one can deny (6) in virtue of a belief in the coherence of the conjunction of the can-claim and the backtracker, and thus one can deny (6) even while accepting Van Inwagen's claim that no one can so act that a law of nature which does obtain

wouldn't obtain. This claim has it that if doing X requires the violation of a natural law, then one *can't* do X. But one can *agree* with this claim and still hold that an agent might be able to render a proposition about a natural law L false. This is because (as I have just shown) one might believe that an agent can render L false *without* being able to perform an act which is such that, were he to perform it, L wouldn't obtain.

Now it might be thought that Van Inwagen has already offered reason to think that the conjunction of a can-claim and a backtracker is incoherent insofar as he has defended premise (5). But this is a mistake, since all that he has established (trivially, I might add) is that one cannot (on account II) render a proposition about the past false. But it clearly *doesn't follow* that one cannot perform some act which is such that, if one were to perform it, the past would have been different from what it actually was. As far as I can see, then, Van Inwagen's defence of (6) is inadequate: it doesn't address itself at all to a possible way of denying (6).

But is it reasonable to think that a conjunction of a can-claim and the relevant backtracking conditional is coherent? It is important to see that this view does *not* commit one to the claim that one can initiate a backward-flowing chain of causation; the relevant 'if' has to be non-causal, and the claim is that there is no reason to suppose that the truth of a "backtracking" conditional rules out the truth of the pertinent claim about freedom. So, it is important to distinguish the following two principles which express different conceptions of the "fixity" of the past:

> (FP1) If $e_1$ occurred at $t_1$, then no agent can at any time later than $t_1$ initiate a causal sequence issuing in $e_1$'s *not* occurring at $t_1$.
>
> (FP2) If $e_1$ occurred at $t_1$, then no agent can at any time later than $t_1$ perform an action such that if he were to perform it, $e_1$ would not have occurred at $t_1$.

Whereas it is quite clear that (FP1) is valid, it is *not* so evident that (FP2) (with the 'if' interpreted non-causally) is also valid. One might mistakenly think that (FP2) is valid by failing to distinguish it from (FP1).[4] And the truth of premise (6) requires not only the truth of (FP1), but also (as I have argued above) that of (FP2).

Van Inwagen might insist that he would reject even (FP2) and that this rejection is *not* based on any sort of confusion. He might employ the following kind of evidence to support his position:

> Consider, for example, the propositions
>     The Spanish Armada was defeated in 1588
> and
>     Peter Van Inwagen never visits Alaska.
> For all I know, the conjunction of these two propositions is true. At any rate, let us assume it is true. Given that it is true, it seems quite clear that I can render it false if and only if I can visit Alaska. (pp. 72–73)

But this sort of evidence is, I think, inconclusive. It might well be true that no one can now perform an act which is such that, were he to perform it, the Spanish

---

[4] For this point, see: Richard Foley, "Compatibilism and Control Over the Past", *Analysis* 39 (1979), pp. 70–4; John Martin Fischer, "Incompatibilism", *Philosophical Studies* 43 (1983), pp. 130–1.

Armada *wouldn't* have been defeated in 1588. But it *doesn't* follow that no agent can ever perform an act which is such that, were he to perform it, the past would have been different (in *some* respect) from what it actually was. The compatibilist wants to say that it is possible that a certain sort of *conjunction* obtain – a conjunction of a can-claim (a claim specifying that an agent can perform some act X) and a relevant backtracking conditional (a claim saying that if the agent were to do X, the past would have been different in some respect from what it actually was). Now it seems to be true that we intuitively would never assent to the conjunction 'Agent A can do X and if he were to do X, then the Spanish Armada wouldn't have been defeated in 1588'. But this might not be because we never would assent to *any* conjunction of a can-claim and a backtracker. Rather, it might be because there is no available act of which we could truly say that if an agent were now to perform it, the Spanish Armada wouldn't have been defeated in 1588. That is, the plausibility of denying Van Inwagen's conjunction might come from scepticism about the truth of this particular conditional, rather than a *general* view that one can never have a true conjunction of a can-claim and a relevant backtracking conditional. So the example does *not* establish that (FP2) is true and thus does not support premise (6).

As I pointed out above, Van Inwagen doesn't claim that employing account II is the *only* way to avoid the counter-examples to account I. And there is a relatively simple alternative which seems to me to avoid the problems with account I. Consider account III of 'S can render P false':

> It is within S's power to arrange or modify the concrete objects that constitute his environment in some way such that if he were so to arrange them, P would be false.

This differs from I by replacing entailment by the weaker subjunctive conditional. Remember that account I had the unhappy consequence that, in the situation described above, Nasser couldn't render false the proposition that all Nostradamus's predictions are correct. But account III avoids this result. If Nasser had destroyed the Sphinx, then Nostradamus would still have predicted that the Sphinx wouldn't be destroyed. Thus, Nasser would have acted in such a way that Nostradamus's prediction would be false, and hence, on account III, Nasser would have rendered false the proposition 'All of Nostradamus's predictions are true'.[5]

It is important to notice that on account III, premise (5) becomes questionable (for the reasons adduced above). That is, no argument has been given that one can't so act that some fact about the past would have been different, where this does not require backwards causation. At least it is clear that Van Inwagen would need another argument for (5), since his argument *rests on* account II. (p. 73) In order to avoid a certain sort of counter-example to account I, Van Inwagen modified I in such a way

---

[5] It might be objected that on account III, one can render false any proposition which is in fact false. So I can render it false that Walter Mondale is President in 1984. This does indicate (perhaps) that there is a divergence between the pre-analytic notion of being able to render a proposition false (if there is such a notion pre-analytically) and account III. Of course, the argument needn't use a phrase which corresponds exactly to any intuitive pre-analytic idea. And Van Inwagen points out that on his own account II, one can render false any false proposition about the past. So account III appears to fare no worse than Van Inwagen's II in capturing the intuitive idea.

that premise (5) became trivially true; but I have claimed that one *needn't* so modify account I, in order to avoid the counter-examples, and if one adopts account III, premise (5) becomes vulnerable. So there is no account on which both (5) and (6) are uncontroversially true.

# III

## THE FIXITY OF THE LAWS

In section II, I argued that Van Inwagen's defence of premise (6) is insufficient insofar as one can accept all that he says about freedom and laws and still deny (6). In this section, I shall argue that his examples do *not* establish his claim about the relationship between freedom and laws, but only a weaker claim. Thus, there seems to be another way which is open to a compatibilist to reject premise (6).[6] Let us distinguish between two different conceptions of the fixity of the natural laws:

> (FL1)  No agent can ever perform an act which itself would be or would cause a law-breaking event, and
>
> (FL2)  No agent can ever so act that a law-breaking event would (at some point) have occurred.

And whereas it is quite clear that (FL1) is valid, it is not so evident that (FL2) is. Consider again Van Inwagen's example of the engineer and the bureaucrat: the problem is that it does not support the stronger (FL2) over the weaker (FL1). It is intuitively apparent that I cannot fly faster than the speed of light – to do so would be to perform an act which would itself be a law-breaking event. (I suppose that even the attribution of such a power to Superman was an exaggeration for effect!) And no engineer can build a spaceship which travels faster than light. To do so would be to perform an act which would cause law-breaking events. We can easily think of many examples where the fact that a possible action would itself be or cause a law-breaking event impels us to think that a person *can't* perform the action.

But suppose that I fail to raise my hand at time T and determinism obtains. Then it might be true that if I had raised my hand at T, some law which actually obtained *wouldn't have* obtained (perhaps just prior to T). Now, if this were the case, then raising my hand wouldn't itself be or cause a law-breaking event – I do not envisage that I would raise my hand at a speed which is faster than light! All that's true is that if I had raised my hand at T, some law which actually obtained wouldn't have obtained. And it's not so obvious that this fact rules it out that I can raise my hand at T.

The point is that there are two different ways in which it might be true that a person can so act that a natural law would be false. And whereas it is evident that no agent can so act that a law would be false in the first way – by performing an act which would be or cause a law-breaking event – it is not so clear that no agent can so act that a law would be false in the second way – by performing an act which itself would not be (and wouldn't cause) a law-breaking event. At least it is useful to note that the examples typically adduced (and adduced by Van Inwagen) do not themselves

[6] Here I am indebted to David Lewis, "Are We Free to Break the Laws?", *Theoria* 47 (1981), pp. 113–21.

establish the claim that the laws are "fixed", interpreted in the second way. It is *this* interpretation which is required by premise (6) and thus incompatibilism.

Having distinguished two ways of so acting that a natural law would be false, it is useful to distinguish *three* different ways in which one might think that an agent can render a law L false (on Van Inwagen's account II). In the first way, one believes that an agent can act in such a way that the past would have been different from what it actually was and L would have still obtained. (This denies FP2.) In the second way, one believes that an agent can perform an act (which itself wouldn't be or cause a law-breaking event) such that, if he were to perform it, L wouldn't have obtained. (This denies FL2.) And finally, in the third way, one believes that an agent can perform an act which itself would be or cause a law-breaking event. (This denies FL1.) Distinguishing these three ways of rendering L false can illuminate a related argument presented by Van Inwagen:

> Let us use 'T' to designate the moment of time that occurred one half hour ago. Let 'P' designate the proposition that I did not visit Arcturus at T. Let '$P_0$' designate the proposition that expresses the state of the world one minute before my birth. (Note that $P_0$ entails the proposition that at that moment Arcturus and I were separated by a distance of about $3.6 \times 10^{17}$ metres.) Let 'L' designate the proposition that nothing travels faster than $3 \times 10^8$ metres per second. We may now argue:
>
> (1) The conjunction of $P_0$ and L entails P.
> (2) It is not possible that I have visited Arcturus at T and P be true.
> (3) If (2) is true, then if I could have visited Arcturus at T, I could have rendered P false.
> (4) If I could have rendered P false, and if the conjunction of $P_0$ and L entails P, then I could have rendered the conjunction of $P_0$ and L false.
> (5) If I could have rendered the conjunction of $P_0$ and L false, then I could have rendered L false.
> (6) I could not have rendered L false.
> So (7) I could not have visited Arcturus at T.
>
> This seems to be a perfectly cogent and unexceptionable argument for the conclusion that I could not have visited Arcturus at T. Anyone who thinks he can demonstrate that one of the premises of the First Formal Argument is false, must either show that his argument does not also 'demonstrate' the falsity of the corresponding premise of the 'Arcturus' argument, or else he must accept this conclusion and explain why the apparent truth of the premise of the 'Arcturus' argument is only apparent. Perhaps someone will be able to do one of these things, but this project does not look very promising to me. (pp. 76–77)

Having made the distinction above, we can undertake the "unpromising project" as follows. In the "Arcturus" argument, the only way I can render L false is to travel faster then $3 \times 10^8$ metres per second. That is, the only way I can render L false is in the *third* way – by performing an act which *itself* would be a law-breaking event. Thus, one can *agree* with Van Inwagen that the "Arcturus" argument is sound *without*

committing oneself to the soundness of the First Formal Argument, insofar as the latter merely requires that S be able to render L false in one of the first two ways, which have *not* been shown to be incoherent by Van Inwagen! Thus, one can say that one of the premises of the First Formal Argument is false without having to say that the corresponding premise of the "Arcturus" argument is false: this premise is (6).

IV

CONCLUSION

Finally, I wish to consider the kind of response to the arguments I have made above which might be given. Remember, I have not attempted to prove compatibilism; rather, I have been focusing on what I take to be *gaps* in Van Inwagen's argument for incompatibilism. He might respond as follows.[7] "You are simply pointing out that *if* compatibilism is true, then one of the premises of my argument must be false. But this is obvious, since the argument is valid! And it is clearly unfair to *assume* the truth of compatibilism, insofar as this is, after all, the question at issue. That is, you are not entitled to assert the can-claims, given the hypothesis of determinism. Finally, I am much more confident of the truth of all of my premises (interpreted so as to make the argument valid) than I am of compatibilism, which is, at best, a highly controversial philosophical thesis."

I believe that this response seriously distorts the dialectical situation. My position does *not* rest on the assumption that compatibilism is true or that it is intuitively more plausible than all of the premises. Rather, the situation is as follows. Van Inwagen has presented an argument which purports to *establish* incompatibilism. Premise (6) is highly controversial and all of the examples which he adduces are compatible with its being the case that only *weaker* premises than (6) are true; so Van Inwagen does not *yet* have a convincing argument for incompatibilism. If a group of examples supports a principle P2 as firmly as P1, but only P1 renders a given argument valid, then simply in virtue of adducing the examples, one hasn't *yet* decisively supported the argument. Finally, I am not *asserting* the pertinent can-claims. All I am pointing out is that if compatibilism is true, then one must accept the coherence of the conjunction of a can-claim with a certain sort of statement. Of course, I have not *shown* that this sort of conjunction is coherent; it is Van Inwagen's job to show that this sort of conjunction is incoherent, insofar as he purports to establish incompatibilism.

It might be objected that it is unreasonable to demand that a philosopher provide examples which absolutely *require* one to accept his position. It seems that it is *necessarily* the case that if a philosopher argues for a certain general principle by giving examples, a weaker principle can be found that might be the strongest principle the examples support.[8] I believe that it is nevertheless important to see exactly where the gaps are in the incompatibilist's argument and thus to see that it is *not* an argument proceeding from premises that no sane person could doubt, by reasoning no sane person could question, to the conclusion that determinism rules out freedom. Also, it is useful to see that the plausibility of the incompatibilist's argument rests on the claims that if one accepts (FP1) one ought also to accept (FP2) and that if one accepts

---

[7] See pp. 74–5, and "Reply to Narveson", *Philosophical Studies* 32 (1977), pp. 94–6.
[8] I am indebted here to personal correspondence with Peter Van Inwagen.

(FL1) one ought also to accept (FL2). The challenge for the incompatibilist consists in explaining why we ought to proceed from the weaker to the stronger principles; without such an explanation, there remains a gap in the argument for incompatibilism.[9]

*Yale University*

[9] I have developed similar criticisms of Van Inwagen's "modal argument" for incompatibilism in: Fischer, "Incompatibilism", *Philosophical Studies* 43 (1983), and "Freedom and Miracles", forthcoming, *Noûs*. For a recent criticism of Van Inwagen's arguments, see: Terence Horgan, "Compatibilism and the Consequence Argument", *Philosophical Studies* 47 (1985), pp. 339-56.