

## ALTERNATIVE POSSIBILITIES AND RESPONSIBILITY

Timothy O'Connor

*Southern Journal of Philosophy*, 31 (1993), 345-372

### I Introduction

Philosophers who maintain that determinism is incompatible with moral responsibility typically do so on the basis of the following two premises:

- (i) A person is morally responsible for what he has done only if he could have done otherwise.
- (ii) A person could have done other than what he in fact did only if determinism is false.

Harry Frankfurt has dubbed the first of these claims "The Principle of Alternate Possibilities" (PAP). Though this principle is widely accepted, Frankfurt<sup>1</sup> has brought to light a range of cases that (to many) appear to provide grounds for rejecting it.

One of his well-known examples concerns a man named Black who wants Jones to perform a certain action:

[Black] waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something other than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do.

As it turns out, Jones decides for his own reasons to perform the desired action, and Black does not intervene in any way with the deliberative process leading to the decision or the carrying out of the action. Jones, we want to say, was responsible for his action and its immediate consequences. (We will assume that the scenario is "normal" in other respects, so that there are not other considerations affecting our evaluation of Jones' responsibility.) And yet it seems that he could not have done otherwise. (The power and intentions of Black ensure that this is so.) So, Frankfurt concludes, (PAP) is seen to be false. If examples of the sort Frankfurt appeals to do show that (PAP) must be rejected, it opens up the possibility of

adopting "semi-compatibilism" - the position that accepts (ii), but denies that causal determinism is incompatible with moral responsibility.

Recently, Peter van Inwagen<sup>2</sup> has argued that the incompatibilist can concede this verdict as regards (PAP) by endorsing three variants of the principle that are both immune to Frankfurt-style counterexamples and capable of yielding the incompatibilist thesis:<sup>3</sup>

(PPA) A person is morally responsible for failing to perform a given act only if he could have performed that act.

(PPP1) A person is morally responsible for a certain event-particular only if he could have prevented it.

(PPP2) A person is morally responsible for a certain state of affairs only if (that state of affairs obtains and) he could have prevented it from obtaining.

It will be noticed that (PPP1) and (PPP2) are closely related, the difference lying in the ontological status given to the consequences of actions for which we hold persons responsible. I will not be discussing van Inwagen's defense of (PPP1) here, for the following reasons. First, the whole matter of individuating "concrete events" construed as particulars is a difficult one, with very little consensus having been achieved thus far. The problem is compounded in relation to van Inwagen's principle, since evaluating his defense of it requires the specification of essential properties of events, so as to track their identity in counterfactual situations. Second, and more importantly, we may legitimately sidestep this difficult task since, it seems to me, we do not hold persons responsible for the specific events that were caused by their actions, but, rather, for various (more general) states of affairs that obtained in virtue of these events. (When we evaluate past actions, morally or prudentially, we take into account alternative states of affairs that the agent might have brought about instead. And this seems appropriate, given that the propositional content of our (prior) intentions, plans, beliefs, and desires refer not to consequences that are particulars, but rather to states of affairs (of varying levels of generality).)

My principal aim in what follows is to assess the adequacy of the remaining two principles van Inwagen suggests, (PPA) and (PPP2). Van Inwagen's defense of these principles, particularly the latter, has come in for criticism by several recent authors. I will

attempt to show that these criticisms are not cogent. I will also bring out a further, neglected consideration that lends support to acceptance of alternative-responsibility principles of the sort that van Inwagen has proposed.

## **II Heinaman's argument against (PPP2)**

I begin with an examination of van Inwagen's primary principle, (PPP2). In what follows, I first sketch his own initial defense of this principle in the face of Frankfurt-type cases, and then proceed to discuss recent objections that have been brought against it. I argue that these objections fail and, indeed, that one ought to maintain (the intuitively highly plausible) (PPP2).

I begin with a word or two concerning the concept of a state of affairs as I will be understanding it here. States of affairs are universals that may obtain in the world as a result of any of a variety of specific circumstances, or "arrangements of concrete particulars" (as van Inwagen puts it). For example, Susan's car's being driven on Tuesday would obtain if Susan drove her car at some particular time on Tuesday morning, but also if she did so later in the day, or if someone else were to drive her car that day. Van Inwagen employs the following convention for referring to states of affairs: we prefix the letter 'C' (an abbreviation for 'its being the case that') to "eternal" sentences. Hence, we represent the example just cited as 'C(Susan's car is driven on August 25th, 1992)', where the latter date gives an unambiguous denotation of the day we referred to simply as Tuesday.

The discussion of (PPP2) in the literature has centered around the following "Frankfurt counter-example", offered by van Inwagen. We suppose there is a man named Gunnar, who

shoots Ridley (intentionally), an action sufficient for the obtaining of Ridley's being dead, a certain state of affairs. But there is some factor, F, which (i) played no causal role in Ridley's death, and (ii) would have caused Ridley's death if Gunnar had not shot him (or had not decided to shoot him), and (iii) is such that Gunnar could not have prevented it from causing Ridley's death except by killing (or deciding to kill) Ridley himself. (1983, p.172)

Factor F might, e.g., be another agent who intends to shoot Ridley if Gunnar does not or a device implanted in Gunnar's brain by another agent that will cause Gunnar to shoot Ridley if he does not choose to do so on his own. In such a scenario, is Gunnar responsible for the state of affairs, C(Ridley dies)? If so, we would appear to have a counterexample to (PPP2), since it seems that Gunnar is unable to prevent its obtaining.

Van Inwagen, however, thinks he can show that Gunnar is not responsible for C(Ridley dies). He suggests that the most plausible basis one might have for supposing he is responsible is the fact that Gunnar did something that was sufficient for that state of affairs, fully intending of his action that it bring about such a consequence. But, he argues, that cannot be a sufficient basis for the attribution of responsibility. For consider the state of affairs C(Ridley is mortal). Gunnar's action seems equally sufficient for this state of affairs, but we clearly do not hold Gunnar responsible for that. Furthermore, we might well suppose that these two states of affairs are in fact equivalent. For the eternal sentences contained in each of them seem to express the same proposition, one which could equally well be expressed by 'Ridley dies at some time or other'. (pp.172-3)

This maneuver will not do, however, as Robert Heinaman has convincingly argued.<sup>4</sup> For C(Ridley is mortal), far from being equivalent to C(Ridley dies), is rather a necessary condition for its obtaining, in just the way that C(The glass is brittle) is a precondition on C(The glass is shattered). But in shattering the glass with a hammer, I do not thereby bring it about that the glass is brittle. Interpreting the notion of 'sufficiency' in the suggestion van Inwagen disputes as causal sufficiency yields a principle that is not obviously incorrect:

(\*) An agent is responsible for a state of affairs if he knowingly and intentionally performed an action that, given the circumstances, was causally sufficient for the obtaining of that state of affairs. (p.271)<sup>5</sup>

And, Heinaman would argue, Gunnar's shooting Ridley was causally sufficient for the obtaining of C(Ridley dies), but not of C(Ridley is mortal).<sup>6</sup>

Heinaman goes on to give an intricate critical discussion of another argument of van Inwagen's for the conclusion that Gunnar is not responsible for C(Ridley is killed). (If the argument is successful, it would point the way to achieving the same result for C(Ridley dies).) The argument may be summarized as follows. Gunnar is not responsible for the state of affairs

(K) C(Ridley is killed by someone who is caused to kill him by F or Ridley is killed by someone who is not caused to kill him by F)

But (K) is equivalent to C(Ridley is killed).<sup>7</sup> Hence, Gunnar is not responsible for the latter state of affairs.

Van Inwagen bases his claim that Gunnar is not responsible for (K) on his earlier attempt to show that the fact that an agent performed an act sufficient for the obtaining of some state of affairs does not entail that he is responsible for its obtaining (i.e., that principle (\*) is false). As we have seen, however, Heinaman has effectively shown that this argument is flawed. I think a strong case can be made for rejecting (\*), though, on the basis of another example van Inwagen introduces near the end of his discussion (pp.176f.): Suppose that Ryder's horse Dobbin has run away with him. Ryder is unable to stop the horse (or to get off), but he can determine the path they will take through the use of his bridle. Ryder directs Dobbin on a course that he knows leads to Rome, in the hope that the runaway horse will result in the injury of some of its citizens, whom Ryder detests. Unbeknownst to him, however, all paths lead to Rome. So no matter what Ryder had done, he could not have prevented C(Ryder passes through Rome on a runaway horse). Ryder's actions are causally sufficient in the circumstances for the obtaining of this state of affairs (as (\*) requires), but it seems clear that he is not responsible for it, since it would have obtained no matter what he had done.

It seems to me that this example is effective in refuting (\*).<sup>8</sup> But if so, van Inwagen contends, we have no equally plausible principle on the basis of which to contest the verdict of (PPP2) that Gunnar is not responsible for (K). And since this is simply a disjunctive

elaboration of C(Ridley is killed), he is not responsible for this state of affairs either. (A more direct argument for this claim is that in the absence of a principle such as (\*), there seems to be no difference relevant to Gunnar's responsibility between (K) and

(M) C(Ridley is killed by Gunnar or grass is green).

Gunnar is clearly not responsible for (M), so it must also be the case that Gunnar is not responsible for (K).<sup>9</sup>

Now Heinaman's strategy is to develop a similar argument for the conclusion that Gunnar is responsible for C(Ridley is killed), and then proceed to show that considerations developed in the argument undermine the support to which van Inwagen's contrary argument appealed. Suppose, then, that we have a case in which factor F is another agent poised to kill Ridley if Gunnar does not. It seems that Gunnar clearly is responsible for C(Ridley is killed by Gunnar). For Gunnar could have prevented this particular state of affairs simply by not shooting Ridley. Now consider the following plausible principle:

(I) If S is responsible for C(A), then if 'F' is any false statement (other than '¬A') then S is responsible for C(A or F).

Since Gunnar is responsible for C(Ridley is killed by Gunnar), it follows from (I) that Gunnar is also responsible for

(L) C(Ridley is killed by Gunnar or Ridley is killed by someone other than Gunnar)

But just as (K) can be seen to be equivalent to C(Ridley is killed), the same appears to hold true of (L). So we arrive at the conclusion that Gunnar is responsible for C(Ridley is killed).  
(p.272)

But notice that if the above argument is accepted, we may readily dispose of van Inwagen's argument for the opposite conclusion. For just as he must concede that in the imagined scenario Gunnar is responsible for C(Ridley is killed by Gunnar), it seems equally clear that Gunnar is responsible for C(Ridley is killed by someone who is not caused to kill him by F). For if he had refrained from shooting Ridley, this state of affairs

would not have obtained. But this, together with the fact that it is false that Ridley was killed by someone who is caused to kill him by F, entails by principle (I) that Gunnar is responsible for

(K) C(Ridley is killed by someone who is caused to kill him by F or Ridley is killed by someone who is not caused to kill him by F)

Briefly stated, then, Heinaman's rejoinder is that there

does seem to be a relevant difference between [M] and [K]: in [K] the second disjunct is false, whereas in [M] the second disjunct is true. Hence, principle (I) allows us to conclude that Gunnar is responsible for [K], while it does not allow us to conclude that Gunnar is responsible for [M]. (p.273)

### III Reply to Heinaman

It is evident that Heinaman's case crucially depends on the truth of (I). But why should we suppose that (I) is true? In answering this question, we should begin by noticing that Heinaman has not succeeded in properly formulating the principle he wants here. The parenthetical clause "other than  $\neg A$ " is intended to avoid attributing responsibility to an agent in the special case in which the disjunction "A or F" is a necessary truth. But it is trivially the case that  $\neg A$  is not the only substitution for 'F' that yields this result. For any expression that is logically entailed by  $\neg A$  obviously does the trick as well. So Heinaman should replace the exclusionary clause with "such that 'A or F' is not a logically necessary truth".

However, the resulting principle is in need of further, more substantive reformulation. For if 'B' is a true proposition depicting some event that occurred at a time prior to the occurrence of the event making 'A' true, and for which S was not responsible, then neither is S responsible for 'A or (B and  $\neg A$ )', and yet this is not a necessary truth and its second disjunct is false. ('A or (B and  $\neg A$ )' is logically equivalent to 'A or B', which was made true when 'B' was.<sup>10</sup>) In the light of this example, it seems to me that what Heinaman needs is the following:

(I') If S is responsible for a state of affairs A, then if 'F' is any false statement (and such that 'A or F' is made true when A is)<sup>11</sup> then S is responsible for the state of affairs that A or F.

Having made the necessary revisions, let us return to the question of the plausibility of (I'). Van Inwagen asks us (p.174) to consider the principle

(\*\*) If a certain state of affairs would have obtained no matter what x had done, then x is not responsible for it.

He notes that the principle seems highly plausible, and there are no features of the sort of case we have been considering that should clearly lead us to reject it. Furthermore, it can be seen to imply the falsity of (I'), for it has the result that Gunnar is not responsible for (L) C(Ridley is killed by Gunnar or Ridley is killed by someone other than Gunnar)

in the case above, whereas (I') (as Heinaman shows) entails that Gunnar is so responsible.

However, as van Inwagen notes, (\*\*) is not completely capable of yielding all the results he wants. Suppose that factor F does not consist of another agent waiting in the wings to shoot Ridley if Gunnar does not, but rather of a mechanism attached to Gunnar's brain such that if he decides not to shoot Ridley, the mechanism will cause the decision to be reversed so that he shoots him nonetheless. In such a case, the antecedent of (\*\*) will not be satisfied, since it is not the case that (L) would have obtained no matter what Gunnar had done. If Gunnar had not shot Ridley, (L) would not have obtained. (It's just that his doing so is rendered inevitable.) But van Inwagen suggests a similarly intuitive variant of (\*\*) that accommodates the case at hand:

(\*\*\*) If a certain state of affairs would have obtained no matter what choices or decisions x had made, then x is not responsible for it.

He comments that

this principle seems at least as evident as the 'no matter what he had done' principle, and it obviously entails that, in the revised case, Gunnar is not responsible for [(L)]. Moreover, this second principle could be applied in the third-party case [our original scenario]... (p.174)

Heinaman attempts to rebut this approach to defending (PPP2) in the following way. We may consider a scenario like the ones we have been discussing, except that it is devoid of any of the Frankfurt-style counterfactual trappings, i.e., Gunnar freely shoots Ridley, and Ridley would not have been killed had Gunnar not done so. (Call this latter situation 'Case 2', and the original 'Case 1'.) Gunnar, of course, is thereby responsible for the state of affairs

(L) C(Ridley is killed by Gunnar or Ridley is killed by someone other than Gunnar)

Heinaman thinks that an attempt to show why Gunnar is responsible for (L) in Case 2 must appeal to (I'): He is responsible for the first disjunct and hence, by (I'), for the disjunction. For (L) obtains in both cases, and what makes it true is the same for each - the truth of the first disjunct. Therefore,

what it is for L to obtain cannot involve anything in Case 1 that is not present in Case 2. (p.274)

Now in the second case the states of affairs

(J) C(If Ridley had not been killed by Gunnar then he would have been killed by someone else)

and

(H) C(It was unavoidable that one or the other of the disjuncts in (L) would obtain)

do not obtain. And so from what has already been said it follows that (L)'s obtaining in Case 1 is independent of both of them. Heinaman sums up his argument as follows:

Obviously, in Case 1 Gunnar is not responsible for the fact that J obtains or for the fact that H obtains. But once J and H are clearly distinguished from L, Gunnar's lack of responsibility for J or for H can be seen to provide no support for the claim that Gunnar is not responsible for L. (p.275)

So far as I can understand Heinaman's remarks here, he is making some sort of direct inference from conditions in virtue of which a state of affairs obtains to conditions sufficient for the ascription of responsibility to agents. (He may, in fact, simply be running together these two sorts of conditions.) That is, his argument seems intended to establish first (what surely is wholly uncontroversial) that (L) obtains as a direct result of Gunnar's action, and

from this he takes himself to be entitled to conclude that Gunnar is responsible for (L). But this relies upon (\*), which I have already criticized. I am somewhat uneasy about reading the argument this way, since it is unclear why Heinaman should regard it as necessary to argue for the claim about what makes (L) true. If, instead, when Heinaman asserts that "what it is for L to obtain cannot involve anything in Case 1 that is not present in Case 2", he means that whatever is responsible for (L)'s obtaining cannot involve anything in Case 1 that is not present in Case 2, then he is simply begging the question against van Inwagen. We need not appeal to (I') to explain Gunnar's responsibility for (L) in the normal scenario of Case 2. Rather, Gunnar is responsible for (L) since he freely performed an act which was sufficient (in the circumstances) for its obtaining, and it would not have obtained if he had not chosen to do so. (This latter clause of course satisfies the necessary condition on responsibility prescribed by (\*\*\*)). I conclude, therefore, that Heinaman has failed to overturn van Inwagen's intuitively appealing defense of (PPP2) via (\*\*) and (\*\*\*)<sup>12</sup>

#### **IV General remarks on the reply to Heinaman**

It might be objected that (\*\*) and (\*\*\*) cannot properly be used in an argument for (PPP2), since they seem to presuppose its truth. But van Inwagen shows that a careful comparison of these principles reveals that neither (\*\*) nor (\*\*\*) entails (PPP2), and hence they cannot be said to presuppose its truth. To enable us to see this more clearly, let us set them out once again, contraposing (PPP2):

(\*\*) If a certain state of affairs would have obtained no matter what x had done, then x is not responsible for it.

(\*\*\*) If a certain state of affairs would have obtained no matter what choices or decisions x had made, then x is not responsible for it.

(PPP2) If x could not have prevented a certain state of affairs from obtaining, then x is not responsible for it.

We may schematically represent them as  $(p \rightarrow q)$ ,  $(r \rightarrow q)$ , and  $(s \rightarrow q)$ , respectively.<sup>13</sup> Van Inwagen employs the fact that (where p, q, and s are contingent) if  $(p \rightarrow q)$  entails  $(s \rightarrow q)$ , then s must entail p. So if either (\*\*) or (\*\*\*) entails (PPP2), then

'x could not have prevented a certain state of affairs from obtaining' must entail 'a certain state of affairs would have obtained no matter what x had done' or 'a certain state of affairs would have obtained no matter what choices or decisions x had made'. But, as he shows through an example, this is clearly not the case. Consider an individual A who cannot refrain from drinking when drink is available to him. If drink is in fact given to him at time t, then he cannot prevent the obtaining of C(A drinks at a time shortly after t). But it is not the case that this would have obtained no matter what he had done, or no matter what decisions he had made. If he had managed to choose not to drink, this state of affairs would not have obtained. So it seems that neither (\*\*) nor (\*\*\*) "presuppose" (PPP2), in the sense of entailing it.

One might argue that van Inwagen is able to achieve this result with respect to (\*\*\*) only through a certain inexplicitness in its formulation. That is, when (\*\*\*) is made fully explicit, it is no longer clear that it does not entail (PPP2). For consider again the scenario which led us to formulate (\*\*\*). There is a mechanism attached to Gunnar's brain which monitors his thought processes and is such that if he decides not to shoot Ridley, the mechanism will be activated in such a way as to cause the decision to be reversed - it will cause a subsequent decision to shoot Ridley. Now one might claim that, as stated,

(\*\*\*) If a certain state of affairs would have obtained no matter what choices or decisions x had made, then x is not responsible for it.

does not yield the result that Gunnar is not responsible for shooting Ridley. For suppose that Gunnar had originally decided not to shoot Ridley. Then, we are supposing, the mechanism would have been activated and caused a subsequent decision to shoot Ridley. But this latter decision would also have been a decision of Gunnar's (albeit one caused by an external factor), and it is not true that if that decision had not been made, C(Ridley is killed) would have obtained nonetheless.

If this objection is cogent, then for (\*\*\*) to function in the way van Inwagen intends, it looks as if we will have to reformulate it as

(\*\*\*^) If a certain state of affairs would have obtained no matter what choices or decisions x had made that x could have made, then x is not responsible for it.<sup>14</sup>

It was not open to Gunnar to decide not to shoot Ridley and maintain that decision in such a way that C(Ridley is killed) would not have occurred. Such a decision sequence is not one that Gunnar could have made.

But if we reconsider van Inwagen's argument that (\*\*\*) does not entail (PPP2), it seems that his example will no longer do the job when (\*\*\*) is replaced by (\*\*\*)^). For we now need to show that

(1) x could not have prevented a certain state of affairs from obtaining

does not entail

(2) a certain state of affairs would have obtained no matter what choices or decisions x had made that x could have made.

So we need to be able to describe a scenario in which (1) is true but (2) is false. In van Inwagen's example, an agent A, who (we suppose) is unable to refrain from drinking when drink is available to him, is given something to drink and drinks it. Given the circumstances, (1) is certainly true. But is (2) false? Is a decision not to drink a decision which A could have made? It seems not. Being unable to refrain from some activity certainly involves (if indeed it is not constituted by) the inability to choose not to engage in it. So we need another example to show that (1) does not entail (2). I am unable to come up with one. Moreover, it seems plausible to suppose that there could not be one. How could it be the case that a state of affairs S is such that an agent A could not have prevented it, and yet is also such that if A had made a choice that was open to her, that she could have made, it would not have obtained? Since this seems inconceivable to me (and I feel fairly confident that van Inwagen would agree), I conclude that (\*\*\*)^ does entail (PPP2).

Having accepted that point, I am not at all convinced that van Inwagen must concede that his argument on behalf of (PPP2) "presupposes" its truth. For I don't believe that he

needs to accept the proposed modification of (\*\*\*)). The claim that it does need to be modified was based on the apparent fact that (in the scenario developed) it is not true that C(Ridley is killed) would have obtained no matter what decisions Gunnar had made - had it been the case that (a) he decided not to on his own and (b) a subsequent, contrary decision was not produced by the mechanism, that state of affairs would not have obtained.

But notice that this sort of move assumes that a "decision" registering in Gunnar's brain as a direct result of some external mechanism would be a decision which Gunnar had made. But it seems to me that the concept of a "decision" involves a degree of autonomy on the part of the agent. I would even suggest that in the highly rare cases in which an agent's action is a direct consequence of psychological compulsion, it would be inappropriate to speak of her as "deciding" or "choosing" anything. But it is even more clearly inappropriate, it would seem, to describe a mental event that was somehow directly effected by external neurophysiological stimulation as a decision of the agent. If, then, we reject this assumption in the case of an externally caused decision, there no longer is a basis for claiming that (\*\*\*) is unable to serve van Inwagen's purposes.

No doubt my rejection of the possibility of an externally caused decision is controversial. Moreover, (\*\*\*) is very similar to (PPP2), and so one who is inclined to think that there are strong intuitions in favor of holding Gunnar responsible for C(Ridley is killed), and therefore reason to reject (PPP2), is likely to want to reject (\*\*\*) as well. Of what value, then, is appeal to (\*\*\*) in a defense of (PPP2)? Let us review the dialectic of the argument thus far. (PPP2) itself is an initially highly intuitive principle. But we are confronted with a range of cases designed to challenge it. In these cases, we are strongly inclined to say that the agent bears responsibility for certain consequences that he brings about.

Up to this point, My defense of (PPP2) has been only partial. For I have tried to defend only the negative thesis that Gunnar is not responsible for a particular state of affairs, C(Ridley is killed). But a convincing defense of this thesis, and of (PPP2) itself,

requires our showing that there is another, closely related state of affairs for which we may plausibly hold the agent responsible, and which does not violate (PPP2). If so, we would seem to have a natural solution to the problem that has been posed. As van Inwagen notes, a candidate which seems equal to the task is

(N) C(Ridley is killed by Gunnar on his own<sup>15</sup>).

And indeed, this would seem to point to a general formula applicable to any Frankfurt-type situation for characterizing a state of affairs for which the agent may be held responsible. For in all such cases, the agent is in no way caused to act or decide as he does, but rather acts or decides "on his own" or freely.

Depending on the particular situation, there may be other states of affairs, which are more broadly delineated, for which the agent is equally responsible. But I think we should recognize, quite apart from considerations stemming from Frankfurt-type scenarios, that from the facts that an agent is responsible for a state of affairs S and that S entails S\*, it does not follow that the agent is responsible for S\*. C(Ridley is killed at t), for example, entails C(The universe exists at t). It seems to me quite natural and intuitive to say that the point of "cutoff" in terms of responsibility in a sequence of increasingly less specific states of affairs (where each entails the one subsequent to it) is precisely the point at which a state of affairs is such that the agent could not have prevented it. That we need not absolve the agent of moral responsibility in Frankfurt cases in order to preserve this intuition seems sufficient to hold that it ought to be preserved - that we have yet to be shown any reason for abandoning it.

In order to reinforce this conclusion, I might emphasize that by ascribing responsibility to Gunnar for (N), but not for C(Ridley is killed), we are not in any way diminishing the extent to which his conduct is reprehensible and blameworthy. We are simply recognizing that care needs to be exercised (especially in highly-contrived scenarios such as we have been considering) in determining precisely which of a number of closely related states of

affairs the agent actually brought about by his action, — and for which he is accordingly responsible — relying on the intuitive notion that an agent cannot be responsible for a state of affairs which he could not have prevented from obtaining. (And we obtain the same results if we rely on the similarly intuitive — and slightly weaker — principle (\*\*\*)).

## V Replies to Rowe and Fischer & Ravizza

We have thus seen that i) acceptance of (PPP2) does not commit one to denying that an agent bears responsibility for some state of affairs in cases where ascription of responsibility clearly seems warranted, and ii) holding an agent responsible for one but not the other of a pair of closely related states of affairs in a Frankfurt-style scenario does not entail a diminished degree of moral reprehensibility on the part of the agent, as compared with a similar case (minus the counterfactual set-up) where the agent is responsible for both states of affairs. These two points suffice, I think, to rebut a couple of recent attempts to show that acceptance of (PPP2) leads to an implausible assimilation of cases.

William Rowe<sup>16</sup> asks us to consider the following three cases, where the latter two are variations on the first:

(Case A) Suppose there is a speeding train approaching a fork in the track controlled by a switch. The left fork (No.1) leads on to where a dog has been tied to the track. If the train proceeds on 1 it will hit the dog. Track No.2, however, leads to a safe stopping point for the train. The switch is set for 2. You have it in your power to throw the switch to 1 or to leave it as is. You throw the switch with the result that the train proceeds on 1, hitting the dog.

(Case B) ...Unfortunately, unlike case A, both tracks 1 and 2 converge later at the point where the dog is tied to the track. It is inevitable, therefore, that the train will hit the dog. Nevertheless, you throw the switch so that the train proceeds on track 1.

(Case C) ...[Here] we find a curious mixture of features in one or the other of our first two cases. As in case A, track 2 does not converge with track 1. Instead, it leads to a safe stopping point for the train. Only track 1 leads to the spot where the dog is tied to the track. Unlike case A, however, some other person, Peter, is so situated that he most certainly will throw the switch if, but only if, you do not. If you throw the switch, the train will be routed to track 1 and hit the dog. If you do not throw the switch, Peter will, with the result that the train will be routed to track 1 and hit the dog. Moreover, it is not in your power to prevent Peter's throwing the switch, should you not throw it yourself. As in our other two cases, you throw the switch, the train is routed to track 1 and hits the dog.

I take it that we are to assume that in each of these cases you (the agent) are aware of the dog's situation and of the fact that track 1 leads to his location. It also seems necessary to assume that in Cases B and C, you are not aware of the fact that the dog will be struck even if you don't throw the switch to track 1. Under these conditions, it is uncontroversially true that in Case A, you are responsible for C(the dog is killed). Further, we have a sense that you bear responsibility for what occurs in Case C, a responsibility which is lacking in Case B.<sup>17</sup> Rowe thinks this just is responsibility for C(the dog is killed), and he correctly notes that (PPP2) does not allow us to analyze the difference in this way. But he fails to recognize that this does not commit the proponent of (PPP2) to asserting that Cases B and C are symmetrical with respect to responsibility. In Case C, but not B, you are responsible, e.g., for

(P) C(The dog is killed as a consequence of your free action)

And, furthermore, there is no reason to suppose that you are less morally culpable for your action in Case C than you are in Case A. For purposes of assessing degree of guilt or culpability, C(the dog is killed) and (P) are not relevantly different.

Similar remarks apply to scenarios constructed by John Martin Fischer and Mark Ravizza.<sup>18</sup> In "Missile 1", an agent (Elizabeth) launches a missile toward Washington D.C.. Had she chosen to do otherwise, a device would have caused the reversal of that decision, thereby bringing about her launching the missile. In "Missile 3", Elizabeth has already launched the missile, but there is another agent, Joan, possessing a weapon which, when activated, is capable of deflecting the trajectory of the missile so that it hits a less populous part of the city. (But the circumstances are such that she cannot prevent the bomb from hitting the city at all.) And in fact she does activate the weapon, and the missile is successfully deflected to another part of the city.

Fischer and Ravizza claim that

...a basic problem for [van Inwagen's] approach is that it forces one to say that "Missile 1" ...and "Missile 3" are on a par. But we believe that, whereas in "Missile 3" Joan is not morally responsible for the consequence-universal, that Washington

D.C. is bombed, in "Missile 1" ... Elizabeth is morally responsible for the consequence-universal, that Washington D.C. is bombed. We believe, then, that van Inwagen's approach implies an implausible assimilation of cases.

It is not clear here whether they are claiming that the cases are "on a par" in all respects, according to (PPP2), or only with respect to C(Washington D.C. is bombed). But even if they are only making the latter, unobjectionable claim, we may plausibly deny (as we have seen) that it is the undesirable outcome they take it to be. In "Missile 1", Elizabeth is morally responsible for

C(Washington D.C. is bombed by Elizabeth, as a consequence of her free action)

and this implies no lesser level of responsibility than that resulting from bringing about C(Washington D.C. is bombed), where no "counterfactual device" is involved. Surprisingly, Fischer himself has recognized the need for making this type of distinction in assessing responsibility for failures to act in Frankfurt-style scenarios, such as the one we considered in connection with (PPA). An individual might be no less culpable for failing to try to perform an action that, owing to a Frankfurt setup, he could not have performed than another person (in similar circumstances) who could have performed the action if she had so chosen. As Fischer comments:

Very roughly, what seems crucial to our moral assessment of persons and our practice of praising and blaming is the person's motivation (and his attempt to act on this motivation). Agents who have the same intentions and make the same choice and are equally conscientious in attempting to act on the choice are accessible to the same degree of praise or blame, even if what they are responsible for is different. The [Frankfurt] examples show that the nature and extent of moral praise or blame do not vary in a straightforward way with changes in the specification of what the person is responsible for: degree of praiseworthiness or blameworthiness needn't vary with content of moral responsibility.<sup>19</sup>

## **VI A problem for some alternative approaches**

There is a further, neglected aspect to the question of responsibility for various states of affairs in Frankfurt-style cases that I think is important: the question of what

responsibility the agent "waiting in the wings" may or may not bear for some of the states of affairs that obtain. I will not here attempt to discuss the particular sort of case (of which our original scenario is an instance) in which this agent has merely formed an intention to act if the other agent were not to do so, since this would force us to address the question I earlier raised and set aside, whether counterfactuals describing what the agent would have freely done under certain circumstances are ever true. So let us restrict our attention to the case in which an agent has attached a device to Gunnar's brain which monitors his thought processes and, should he choose not to shoot Ridley, would automatically cause him to change his decision and proceed to do so. (Following van Inwagen's own story, let us refer to this other agent as 'Cossar'.)

Finally, in order to underscore the point I will be making, we may suppose that the mechanism which is attached to Gunnar's brain has not been determined all along to cause Gunnar to change his mind about shooting Ridley if he were on his own to choose not to do so. Rather, it must first be activated by pressing a button on a hand-held remote control device possessed by Cossar. Also, there are no means at Cossar's disposal for de-activating the mechanism once the button has been pushed. Just as Ridley is about to walk on the scene (prior to Gunnar's decision), Cossar activates the mechanism.

It would seem that, in those circumstances, Cossar's action is causally sufficient for the obtaining of

(K) C(Ridley is killed by someone who is caused to kill him by F or Ridley is killed by someone who is not caused to kill him by F)

as well as for

C(Ridley is killed)

(which is equivalent). But then it appears that Heinaman's (\*)

(\*) An agent is responsible for a state of affairs if he knowingly and intentionally performed an action that, given the circumstances, was causally sufficient for the obtaining of that state of affairs.

licenses the attribution of responsibility to Cossar for the states of affairs mentioned. This is clearly a counterintuitive result. (I might also note that Heinaman is committed to accepting the consequence that both Cossar and Gunnar are responsible for (K), since the latter's action also was causally sufficient (given the circumstances) for its obtaining.)

The same problem may afflict the account advanced by Rowe in the article in which his criticism of van Inwagen appears. The account actually is put forth as giving necessary and sufficient conditions on an agent S's causing a state of affairs E by doing X, but we may ignore this complication, since Rowe takes S's causing E to be the central necessary condition on his being responsible for E. (Indeed, he allows that an agent is prima facie responsible for E if he causes E. He is ultima facie responsible for E if, in addition, he is aware of the relevant circumstances, intending that E result by that action, and so forth.) According to Rowe, then, a person S is prima facie morally responsible for a state of affairs E as a result of doing X if and only if

- (1) S does X prior to or at the same time as E's occurrence, and
- (2) S's doing X is part of a sufficient causal condition of E, and
- (3) either S's doing X is necessary for E's occurrence or any other condition that is sufficient (in the circumstances) for E has a part that is actualized only if S does not do X.

The crucial component of this condition is (3). Rowe formulates (3) in an effort to give a theoretical underpinning to his intuition that, among the three train-track scenarios introduced in the previous section, the agent is morally responsible in (C), but not in (B), for C(The dog is killed). The first disjunct of (3) will be satisfied in "normal" (non-Frankfurt) scenarios, and Rowe takes it that the second disjunct will be satisfied in Frankfurt scenarios in which i) the agent's action is not necessary for the obtaining of some state of affairs, but ii) the circumstance in virtue of which the agent's action is not necessary does not play a role in the production of the state of affairs in the actual sequence. In Case (C), Peter (the agent in the wings) does not causally contribute to the sequence of events in virtue of which C(The dog is killed) obtains, and so, apparently, condition (3) is satisfied. (Peter's intention to throw the switch if you do not is the only other sufficient condition for C(The dog is

killed), given an intuitive sense [which I shall not attempt to explicate here] of the notion of genuinely distinct sufficient conditions.)

Now there is a significant unclarity surrounding the omission of a temporal index on "any other condition that is sufficient (in the circumstances) for E" in (3). Let us suppose first that Rowe intends there to be no restriction on the time of any such condition (other than its being prior to or simultaneous with the occurrence of E). If so, it seems to me that the resulting principle ought to be deemed inadequate (because too strong) by those who (unlike me) follow Rowe in rejecting (PPP2) upon reflecting on cases such as (C).

For consider the following variation on Fischer and Ravizza's "Missile" cases. Elizabeth launches a missile toward Washington, D.C. Several hours later, just before this missile is about to strike the city, Susan launches another missile at the same target. In this scenario, Elizabeth's action fails to satisfy (3) relative to the state of affairs C(Washington,D.C. is bombed), and so she is not deemed responsible for it. Why would this be accepted, though, by one who denies (PPP2)? If it is not relevant to an agent's responsibility that she was unable to prevent a certain state of affairs, why is it relevant that another agent initiated a process after the first agent's action is completed that also ensures that the state of affairs will obtain? In so far as I understand the intuitions behind the rejection of (PPP2), they also lead to the rejection of the unrestricted version of (3).

What, then, of a version of (3) that refers only to other sufficient conditions for E obtaining at or prior to the time of S's doing X?<sup>20</sup> In this case, Rowe is stuck with the unwanted outcome that in the case of Cossar and Gunnar (outlined at the beginning of this section), Cossar is (prima facie)<sup>21</sup> responsible for C(Ridley is killed). For his activating the mechanism is part of an antecedent sufficient causal condition of E (thereby satisfying Rowe's (1) and (2)), and there is no independent sufficient condition for E obtaining at the time of his action (which ensures that (3) is satisfied).

Furthermore, it is far from clear that his account also entails (in accordance with Rowe's intention) that Gunnar is so responsible. For let us take a closer look at his condition (3):

(3) Either S's doing X is necessary for E's occurrence or any other condition that is sufficient (in the circumstances) for E has a part that is actualized only if S does not do X.

Rowe will claim that Gunnar's action satisfies (3), in virtue of satisfying the second disjunct. The activated mechanism, which is the only other contemporaneous condition sufficient (in the circumstances) for E, has "a part that is actualized" only if Gunnar does not decide to shoot Ridley. But precisely what is this supposed to involve? All the components of the sufficient condition involving the activated mechanism are, of course, actual. It seems that Rowe has in mind the fact that this sufficient condition does not determine the particular way in which C(Ridley is killed) will come about — it determines that one of two specific processes will obtain, and each of them leads to the same result. Both of these possible processes are thought to be "parts" of the sufficient condition, but one of them will be actualized only if Gunnar decides not to kill Ridley.

But if that is what is meant by a sufficient condition's having an unactualized part, then the problem of unintended results crops up elsewhere. Recall Case B, in which the pair of tracks stemming from the fork in the track reconverge. Rowe rightly maintains that you are not responsible for C(the dog is killed) in virtue of throwing the switch that causes the train to proceed down track 1. But does his condition (3) permit him to say this? Prior to your throwing the switch, the fact that the train is speeding towards the fork in the track is sufficient (given the circumstances that each of the diverging tracks leads to where the dog is tied and that you are unable to stop the train) for the obtaining of the resulting state of affairs. But if having an "unactualized part" is interpreted as above, then the inexorable progression of the train towards the hapless dog seems to satisfy it. For it determines only that one or the other of two specific processes will lead to the killing of the dog, and one of these is actualized only if you do not throw the switch.<sup>22</sup>

I conclude, therefore, that Rowe's account clearly commits him to maintaining that Cossar (the counterfactual intervener) is responsible for C(Ridley is killed) in our original scenario. Further undesired consequences seem to follow as well, the precise nature of which will depend on how we clarify the notion of a sufficient causal condition's having unactualized parts.

Finally, I will briefly examine the sophisticated theory recently sketched by Fischer and Ravizza.<sup>23</sup> Like Rowe, they suggest that we may specify features of the "actual sequence" in Frankfurt scenarios that ground the agent's responsibility independently of whether there was an alternative action available to the agent such that he could have prevented the state of affairs from obtaining by performing it. I believe that strong doubts may be raised about whether we can isolate features of these cases in the manner they suggest, but I shall not raise them here. Instead, I shall again examine the implications of their account for the question of whether Cossar, in our representative case, bears any responsibility for C(Ridley is killed), in addition to (or instead of) Gunnar.

A summary statement of their account is given in the following:

... actual causal control of a consequence is sufficient for moral responsibility for that consequence.... We shall say that an agent has actual causal control of some consequence insofar as it issues from a responsive sequence.

...[We may] distinguish two components of the sequence leading to a consequence. The first component is the mechanism leading to action (bodily movement), and the second component is the process leading from the action to the event in the external world. We shall say that in order for the sequence leading to a consequence to be responsive, both the mechanism leading to the action must be weakly reasons-responsive<sup>24</sup> and the process leading from the action to the consequence must be "sensitive to action".

Suppose that in the actual world an agent S performs some action A via a type of mechanism M, and S's A-ing causes some consequence C via a type of process P. We shall say that the sequence leading to the consequence C is responsive if and only if there exists some action A\* (other than A) such that: (i) there exists some possible scenario in which an M-type mechanism operates, the agent has reason to do A\*, and the agent does A\*; and (ii) if S were to do A\*, others' behavior were held fixed, and a P-type process were to occur, then C would not occur. (1991, pp.272-3)

As Fischer and Ravizza acknowledge, the notions of a "reasons-responsive mechanism" and a "type of process" leading from the action to the obtaining of a

consequence are vague and in need of fuller explication. I will attempt to work with these notions nonetheless, because I think that the applications of them that I will make should be unobjectionable to Fischer and Ravizza.

Consider Cossar, who performs the action of activating the device attached to Gunnar's brain. I claim that C(Ridley is killed) issues from a responsive sequence originating with this action, and so, on Fischer and Ravizza's account, it is a consequence for which Cossar is responsible.

To substantiate this claim, I need to show that there is a possible action that satisfies the two conditions they stipulate. But first, I will deal with what is likely to be the most immediate objection: that the state of affairs C(Ridley is killed) is not caused to obtain by Cossar's action, but rather by Gunnar's. If we were speaking of the event-particular that we might refer to as "Ridley's death", this objection would seem entirely appropriate. For if Gunnar's action is not causally determined, then Ridley's death would not be a causal consequence of Cossar's action. But a state of affairs (or "consequence-universal", in Fischer and Ravizza's terminology) may obtain in a variety of ways, depending on its level of generality. And in the case at hand, C(Ridley is killed) is causally determined to obtain as a result of Cossar's action. The fact that Cossar's action does not (in this special sort of case) also determine the particular events in virtue of which that state of affairs obtained seems irrelevant to the issue of whether he caused (or causally contributed to) that state of affairs.<sup>25</sup>

Can we, then, specify an action of Cossar's meeting the criteria set out by Fischer and Ravizza? Let our action be that of activating the device in an alternative fashion, such that it ensures that Gunnar will not successfully carry out the action of shooting Ridley. The first criterion concerns the deliberative "mechanism" which may be thought to be operative in Cossar's decision to activate the device in Gunnar's brain. Is it such that it would issue in the alternative action we have specified given some possible incentive? If (as Fischer and Ravizza are supposing) subjunctive conditionals specifying free human actions are ever true,

there seems to be no reason whatever for denying that at least one of the specific sort germane to our present question is true.

Secondly, we need to decide whether the following counterfactual is true: If a) Cossar were to activate the device in the alternative way, b) the actions of other agents were held fixed, and c) a process of the sort leading from Cossar's action to the consequence in the actual sequence were to occur, then C(Ridley is killed) would not obtain. And, again, it seems fairly clear that this counterfactual is true. For the hypothesized action by Cossar together with the surrounding circumstances that actually obtained logically entail the counterfactual's consequent. There seems no reason to suppose that the world would differ in this respect if the counterfactual's antecedent were to have obtained instead.<sup>26</sup> Just what sort of process led from Cossar's action to C(Ridley is killed) in the actual sequence? It is not entirely clear what Fischer and Ravizza have in mind, but all we need to suppose is that it is not the case that its conjunction with the other aspects ((a) and (b)) of the counterfactual assumption would obtain only in worlds so far removed from our own that we could not be confident of the truth of the counterfactual, and I cannot see that this is the case. The sketchy remarks that Fischer and Ravizza make about individuating types of processes indicate that they would do so very coarsely, and so parallel cases in each of which the consequence immediately results from a decision by Gunnar would seem to be clearly of the same "type".

I conclude, therefore, that the alternative to (PPP2) given by Fischer and Ravizza has suffered the same fate as Heinaman's and Rowe's. Each implies (inadvertently) that Cossar is responsible for C(Ridley is killed).

Rejecting these approaches in favor of (PPP2) and (\*\*\*) enables us to avoid such problematic outcomes, while still being able to identify related states of affairs for which each of the agents (i.e., Cossar and Gunnar) is responsible. It is plausible to maintain that, given the peculiar nature of the circumstances, neither agent is responsible for bringing about the obtaining of C(Ridley is killed) by his freely chosen action, since neither of them

could have prevented it merely by refraining from that action — for each agent, C(Ridley is killed) is counterfactually independent of his choice (i.e., it would have obtained, even if he were to have chosen otherwise).<sup>27</sup> But Gunnar is responsible for such states of affairs as C(Ridley is killed by Gunnar on his own) and the state of affairs constituting the second disjunct of (K). And as for Cossar, he is clearly responsible for

C(It is causally inevitable at t-1 that C(Ridley is killed) will obtain)

(where t-1 denotes a time after the mechanism has been activated and prior to Gunnar's shooting Ridley). It is hard to understand why we should want to resist such an analysis, once we see that, in addition to incorporating such intuitively compelling principles as (PPP2) and (\*\*\*) , it can also accommodate our impulse to attribute responsibility to agents such as Gunnar in Frankfurt-type cases.

## VII (PPA)

I will now turn to van Inwagen's other principle, (PPA):

(PPA) A person is morally responsible for failing to perform a given act only if he could have performed that act.

Van Inwagen's defense of (PPA) is brief, and most commentators have not directly challenged it. It consists of spelling out an example that appears to follow the basic strategy of Frankfurt-style scenarios (adapted to a case in which an agent freely decided not to perform some act), and noting that it seems implausible to say that the agent in the example is in fact responsible for failing to perform that act. As van Inwagen remarks, it is "notoriously hard to prove a universal negative proposition." Having shown that the principle survives an example that seems to be properly tailored in the usual way of "Frankfurt-style" scenarios, he challenges the critic to construct a successful counterexample. (p.165-6)

The example van Inwagen employs involves a man who observes a robbery taking place outside his home. He decides not to get involved, and so fails to notify the police. Unbeknownst to him, the city's phone system was temporarily disabled, and remained so for several hours. Is the man responsible for failing to call the police? It would seem that he is not, and it further seems likely that the natural inclination to judge the case in that way involves tacit acceptance of (PPA). If the example is modified so that the agent's phone would become inoperative only if he were to decide to make the call (in closer accordance with the usual Frankfurt scenarios), this conclusion intuitively remains unaffected.

While I thus do think that van Inwagen is correct in supposing that (PPA) is invulnerable to counterexample, one might challenge the appropriateness of his particular formulation of the relevant principle concerning unperformed acts. For, more fundamentally, we form judgments concerning a person's responsibility for failing to try to perform certain acts, since whether or not she succeeds in doing so is not completely up to her. Such judgments would be made even if we knew that (unbeknownst to the agent) there were circumstances that would have prevented her from performing the act, even if she had tried. (Van Inwagen seems to acknowledge this, but does not try to formulate a principle concerning "failures to try".) So we ought to consider whether a plausible "alternative possibilities" criterion of responsibility for 'failures to try' can be formulated.<sup>28</sup> A natural candidate is the following:

(PPA\*) A person is morally responsible for failing to try to perform a given act only if he could have tried to perform that act.

However, it would seem that putative counterexamples are ready to hand. We may, e.g., modify the original case involving the nefarious Black, and suppose that Jones is faced with a decision whether to perform an action that Black very much wants not to be done. If Jones shows any sign of choosing to undertake the action, Black will use his special powers to ensure that Jones ends up not choosing to do it. Here it seems that Jones is responsible for

not trying to perform the action, and yet it is (perhaps) the case that he could not have tried to do it. What are we to make of this?

Quite simply, I think this example shows rather that the initial proposal, (PPA\*), is inadequate to handle such cases. This is not to say that the principle is false, but rather that it fails to govern the scenarios in which we are interested here. For it is clear that the agent in our present example bears some moral responsibility connected with his inactivity, but it cannot be for failing to try to perform some action, for he could not have done even that. The relevant features are precisely parallel to those considered in endorsing (PPA), for trying to X might plausibly be said to be a type of action itself, and so is included in the range of (PPA). One might feel somewhat more hesitant (as I have) about retaining (PPA\*) as against (PPA) in the face of the Frankfurt scenario, but I think such uneasiness can be accounted for in terms of the following two reasons: First, cases of one's bearing some sort of responsibility associated with a failure to perform or try to perform some action in which one nonetheless couldn't have so much as tried to perform it seem to be restricted to the more fantastic of the Frankfurt scenarios. By contrast (as van Inwagen's example shows), there no doubt are actual cases in which ascription of responsibility seems appropriate even though one could not have performed the relevant action. Closely connected to this first reason is the fact that we readily see an alternative to holding an agent responsible for an action that (it turns out) he could not have performed - he might bear responsibility for simply failing to try to perform it. But it might seem that the latter sort of responsibility is, so to speak, the end of the road, and since it's clear in our putative counterexample to (PPA\*) that the agent is in some way responsible, one might conclude that, contrary to (PPA\*), it is for failing to try to act (even though he could not have done so).

I believe, however, that the fact that the cases governed by (PPA\*) are (as I have suggested) a subset of those within the range of the highly plausible (PPA) should prompt us not to reject it. What is more, we already have seen an alternative-possibilities principle that will do the work required by our present example: (PPP2). Van Inwagen offered

(PPP2) to account for responsibility for the consequences of our actions in Frankfurt scenarios, but it is not restricted to these. Let us suppose that the observer of the robbery is Jones. Then (PPP2) implies that Jones is not responsible for C(Jones fails to try to call the police), since he would have been caused not to try if he had even begun to consider the possibility. If a friend were to remonstrate with him for his selfish conduct, Jones might upon finding out that he had been monitored by Black protest that he couldn't have done so - indeed, couldn't have so much as tried. But the seemingly proper reply (permitted by (PPP2)) is "Yes, but you didn't know that, and you persisted in your cowardly and uncaring inactivity of your own free will." In our more circumspect technical terminology, we would express this by saying that Jones (knowingly) brought about the state of affairs C(Jones fails to call the police of his own free will), and since he could have prevented this state of affairs from obtaining, he is (prima facie) responsible for it.

### **VIII Does (PPP2) entail (PAP)?**

A final matter that I wish to consider briefly is whether or not van Inwagen's acceptance of (PPP2) commits him to (PAP). Consider the following argument (suggested to me by Carl Ginet): van Inwagen seems to accept<sup>29</sup> the (highly plausible) principle

- (R) A person is morally responsible for an action of his only if he is morally responsible for some state of affairs for which that act was sufficient (in the circumstances).

Moreover, (PPP2) entails (as a restricted version)

- (PPP2\*) A person is morally responsible for a state of affairs for which an act of his was sufficient only if he could have prevented that state of affairs.

Finally, it also seems to be the case that

- (S) A person could have prevented a state of affairs for which his act was sufficient only if he could have done otherwise.

But taken together, these three principles directly entail

(PAP) A person is morally responsible for an action of his only if he could have done otherwise.

I think this argument is sufficient to show that van Inwagen (and any other proponent of (PPP2)) ought to accept (PAP) in some form or other, but it is apt to be misleading owing to the vagueness of the principle (S) employed in the above derivation. More precisely, the acceptability of (S) in the light of Frankfurt-style cases depends on the phrase "he could have done otherwise" being understood in the restricted sense of "he could have acted in a way such that some aspect of his total action or sequence of actions<sup>30</sup> would have been different." For while it seems evident that one cannot prevent a state of affairs that one brings about if one cannot move one's body or at least deliberate in an alternative fashion, this may involve no more than making a different (initial) decision, as in our scenario in which a device is attached to Gunnar's brain. It is not necessary that the agent be able to perform a different type of action. (I take it that van Inwagen assumed that this latter, stronger requirement is made by (PAP), and that this explains his unwillingness to endorse the principle.) So we need to recognize that this phrase also has a "restricted sense" in the version of (PAP) which constitutes the conclusion of the argument.<sup>31</sup>

<sup>1</sup> "Alternative Possibilities and Moral Responsibility", Journal of Philosophy 66 (Dec. 1969), 829-39.

<sup>2</sup> "Ability and Responsibility", Philosophical Review 87 (April 1978), 201-24; and An Essay on Free Will (Oxford: University Press, 1983), Ch.5. Unless indicated otherwise, page references in the text will be to the latter work.

<sup>3</sup> The names given to these principles are acronyms for "the Principle of Possible Action" and "the Principle of Possible Prevention", respectively.

<sup>4</sup> "Incompatibilism Without the Principle of Alternative Possibilities", Australasian Journal of Philosophy 64 (Sept. 1986), 266-76.

<sup>5</sup> Mark Crimmins has pointed out to me that (\*) is ambiguous, since it doesn't make clear whether the agent is supposed to know about the causal sufficiency. I assume that this knowledge is required, since the principle would be much less plausible otherwise.

<sup>6</sup> I suppose that someone might want to resist Heinaman's analogy of C(Ridley is mortal) to C(The glass is brittle) by claiming that to be mortal is not simply to be subject to or capable of death, but further implies that one will die at some time or other. (And this is just what is expressed by C(Ridley dies).) I will not attempt to argue this matter here, for van Inwagen's argument can be evaded by substituting C(Ridley dies at t) - which is uncontroversially not equivalent to C(Ridley is mortal) - for C(Ridley dies).

<sup>7</sup> Norman Kretzmann has pointed out that, in fact, these do not appear to be equivalent, since Ridley could be killed, say, in a car crash, by no one. (He's driving alone, his tire blows, and he crashes into a tree.) I believe that van Inwagen has in mind something like "is killed by someone" when he uses the simpler term "killed". Rather than modifying the quotations that follow, I will simply stipulate that the expression is to be understood in this way in what follows.

<sup>8</sup> I give another reason below for supposing (\*) to be faulty.

<sup>9</sup> Van Inwagen (1978), pp.164-6; compare Heinaman (1986), pp.272-3.

<sup>10</sup> The expression "made true when B was" is here simply a less cumbersome way of expressing "true in virtue of an event or state of affairs obtaining at the same time as the event or state of affairs in virtue of which B is true."

<sup>11</sup> It should be clear that this clause also implies that "A or F" is not a necessary truth.

<sup>12</sup> I might here mention in passing an assumption underlying the present discussion (as well as most others in the literature) of the standard Frankfurt scenarios involving a counterfactual intervener waiting in the wings, an agent who intends to act in a certain way under possible conditions that turn out to be non-actual. The assumption is that, given the agent's firm intention, a counterfactual proposition describing what he would do if these circumstances were to obtain may be true. In particular, it is true if the consequent states that he would act in accordance with that intention. If, however, an agent's free action under normal circumstances is not causally determined, then it is not at all clear that this assumption is tenable. For there will be possible worlds having the same history (including the truth of the counterfactual's antecedent and the agent's intention) until just prior to the agent's action at *t* which differ in terms of which course of action the agent chooses to follow.

Various responses have been made to this problem, but I will not pursue the matter here. For if it is denied that it is determinately true that the agent in the wings in Case 1 would have shot Ridley if Gunnar had not, then there no longer seems to be a basis for saying that C(Ridley is killed) would have obtained no matter what Gunnar had done, and so the scenario cannot provide a counterexample to (PPP2) (or (PAP), for that matter). We may, however, restore the (apparent) problem by changing the case so that there is a causally-determined mechanism which monitors Gunnar's brain, and which will cause him to shoot Ridley if he should choose not to do so on his own. For expository simplicity, I will continue to make use of our original scenario, assuming that the counterfactual proposition 'If Gunnar were to decide not to shoot Ridley, the other agent would do so' is in fact true.

<sup>13</sup> I am indebted to unpublished notes of Carl Ginet's for the particular way in which I reconstruct van Inwagen's argument in this paragraph.

<sup>14</sup> Carl Ginet has suggested the need for this modification in the unpublished discussion cited in the previous footnote. He introduces it in connection with a slightly different example which he develops there, but the difference is irrelevant to the point I am making.

<sup>15</sup> I.e., as a result of a free decision by Gunnar.

<sup>16</sup> "Causing and Being Responsible for What is Inevitable", American Philosophical Quarterly 26 (April 1989), 153-59.

<sup>17</sup> I am not suggesting that you are wholly blameless in case B. Clearly you are acting upon a malevolent intention, and it seems appropriate to condemn your conduct

for just this reason. But this only shows that there is more to be responsible for in the situation described than states of affairs entailing the dog's death.

<sup>18</sup> "Responsibility for Consequences", in J. Coleman and A. Buchanan, eds., *Festschrift for Joel Feinberg*, forthcoming.

<sup>19</sup> "Responsibility and Failure", *Proceedings of the Aristotelian Society* N.S. 86 (1985/86), 251-70; the quotation is found on p.256.

<sup>20</sup> There is some indication that that is Rowe's intent. See, e.g., his principle (Y) on p.156.

<sup>21</sup> As noted above, other conditions pertaining to the agent's knowledge and intention in acting would have to be satisfied in order to drop this qualification. However, there is nothing about this case (or others I discuss immediately below) to suggest that the agent fails to meet these further conditions. Therefore, I will omit this qualification in discussing the other cases.

<sup>22</sup> After setting out the final version of the account (the one that I have discussed here), Rowe returns to Case B and simply states that the sufficient condition constituted by the train's careening down the track together with the relevant circumstances "has no part that is actualized only if you do not throw the switch" (p.156). But I am unable to understand this notion in such a way that this claim and the assertion that Cossar's activation of the mechanism has an unactualized part are both true. I am forced to conclude that Rowe is unwittingly thinking of this notion in different ways when considering these cases.

<sup>23</sup> "Responsibility and Inevitability", *Ethics* 101 (Jan. 1991), 258-78.

<sup>24</sup> "In order to determine whether an actual-sequence mechanism of a certain type is weakly reasons-responsive, one asks whether there exists some possible scenario in which that type of mechanism operates, the agent has reason to do otherwise, and the agent does otherwise (for that reason). That is, we hold fixed the actual type of mechanism, and we ask whether the agent would respond to some possible incentive to do otherwise. If so, then the actually operative mechanism is weakly reasons-responsive." (p.269)

<sup>25</sup> To reinforce this conclusion, consider a scenario in which a nuclear physicist has rigged an explosive device in such a way that the occurrence of either of two causally possible sub-atomic events will cause it to explode. Furthermore, it is physically necessary that one of these two events should occur. It is clear, I believe, that although the physicist did not cause the specific events that led to the detonation, he nonetheless did cause C(the bomb explodes) to obtain.

<sup>26</sup> The reader may be wondering about Gunnar's action of shooting Ridley in the actual sequence. Is that among those actions that are to be "held fixed"? (If so, the sequence stemming from Cossar's action would not meet the criterion of responsiveness.) While Fischer's and Ravizza's remarks do not provide a direct answer to this question, I think they would say that it is not to be held fixed, given their brief comments on the possibility of "simultaneous overdetermination" of a consequence by the actions of more than one agent. They believe that in this sort of scenario, the agents are jointly responsible for the outcome. Accordingly, they suggest that in order to ascertain if any particular agent's action in such a case was part of a responsive sequence, one must "bracket" the actions of the other agents. (p.274, n.18)

And, at any rate, it is simply implausible to maintain that Gunnar's action must be held fixed in testing the responsiveness of the sequence issuing from Cossar's action. If this is not sufficiently evident, consider a case in which Cossar directly (and freely) causes Gunnar to shoot Ridley (instead of merely ensuring that Ridley's death will result). Here it is uncontroversially the case that Cossar is responsible for C(Ridley is killed). To get this result on Fischer's and Ravizza's theory, though, we must suppose that in testing for responsiveness, Gunnar's action of shooting Ridley is not held fixed in determining the alternative sequence.

<sup>27</sup> This is not to say that there is no choice available to Cossar such that if he were to have made it, then the consequence would not have obtained. He could, after all, have chosen in our revised scenario to activate the device differently, so as to ensure that Gunnar would not have shot Ridley. But this is quite different from supposing him merely to have refrained from the action he did in fact perform. If, given the extraordinary powers available to him in this scenario, we are inclined to hold him responsible for not having made sure that C(Ridley is killed) did not obtain, this concerns the quite different matter of responsibility for the consequence of an omission, rather than for having brought about a consequence by one's action.

<sup>28</sup> As Martha Klein notes in Determinism, Blameworthiness, and Deprivation (Oxford: The Clarendon Press, 1990), p.41.

<sup>29</sup> cf. his remarks on p.181 of (1983).

<sup>30</sup> I needn't for present purposes take sides on the issue of how actions are to be individuated.

<sup>31</sup> I thank Carl Ginet, Norman Kretzmann, and Mark Crimmins for detailed comments on earlier versions of this material. I'm also grateful to Eleonore Stump for helpful discussions on these issues. Finally, I wish to thank John Martin Fischer and Mark

Ravizza for kindly sending me relevant unpublished work of theirs from which I have benefited much.