The Philosophical Review, LXXXVII, No. 2 (April 1978).

ABILITY AND RESPONSIBILITY

Peter van Inwagen

I

There was a time when philosophers would debate the relative merits of the doctrines of "liberty" and "necessity," or, as we should say today, debate whether it is more reasonable to believe in free will or in universal causal determinism. As everyone knows, the parties to this debate shared a premise: that free will and universal causal determinism are incompatible. And, as everyone knows, there arose a philosophical tradition—represented by Hobbes, Hume, Jonathan Edwards, Mill, and Moritz Schlick—in which just this premise is denied. Thus the debate between the libertarians and necessitarians was undercut, and most of the debates about free will today are, as they have been for a long time, essentially debates about whether free will and determinism are compatible or incompatible.

But why should anyone care whether we have free will or whether determinism is true? The first part of this question is perhaps easier to answer than the second: we care about free will because we care about moral responsibility, and we are persuaded that we cannot make ascriptions of moral responsibility to agents who lack free will. Recently, however, Harry Frankfurt has denied just this principle, or, at least, a principle that sounds very much like it, which he calls the Principle of Alternate Possibilities.[1] His formulation of the Principle of Alternate Possibilities is

PAP A person is morally responsible for what he has done only if he could have done otherwise. (p. 829)

If Frankfurt has made out a good case for the falsity of PAP (and I think he has), then it would seem that he has undercut the debate between the "compatibilists" and the "incompatibilists" (to use the contemporary jargon) in a way very similar to

[1] "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy,* LXVI (1969), 829-839.

the way in which Hobbes and others undercut the debate between the libertarians and the necessitarians.

Frankfurt supports his contention that PAP is false by means of a certain style of counterexample; I shall call counterexamples in this style, "Frankfurt counterexamples."[2] The following Frankfurt counterexample is due to David Blumenfeld. It is worked out with rather more concrete detail than any of Frankfurt's own counterexamples:

> Suppose that the presence of a certain atmospheric reaction always causes Smith to decide to attack the person nearest to him and to actually do so. Suppose also that he always flushes a deep red when he considers and decides *against* performing an act of violence and that under certain circumstances the atmospheric reaction is triggered by the appearance of just this shade of red. Now imagine that on a day on which circumstances are favorable to the triggering of the reaction, Smith considers whether or not to strike a person with whom he is conversing, decides in favor of it, and forthwith does so.[3]

The general idea behind Frankfurt counterexamples is this. An agent S is in the process of deciding which of n alternative acts $A_1 \ldots, A_k \ldots, A_n$ to perform. He believes (correctly) that he cannot avoid performing some one of these acts. He decides to perform, and, acting on this decision, does perform A_k. But, unknown to him, there were various factors that *would have* prevented him from performing (and perhaps even from deciding to perform) any of $A_1 \ldots, A_n$ *except* A_k. These factors would have "come into play" if he had shown any tendency towards performing (perhaps even towards deciding to perform) any of $A_1 \ldots, A_n$ except A_k. But since he in fact showed no such tendency, these factors remained mere unactualized dispositions of the objects constituting his environment: they played no role whatever in his deciding to perform or in his performing A_k.

According to Frankfurt, it is evident that in such cases we should say (*i*) that S had no alternative to performing A_k, couldn't have done otherwise than perform A_k, and (*ii*) S is nonetheless responsible for having performed A_k (or, at least, if he is *not* responsible for having performed A_k, this must be due to some factor

[2] Or perhaps they should be called "Frankfurt-Nozick" counterexamples. See *n.* 2 (p. 835) to Frankfurt's article.

[3] David Blumenfeld, "The Principle of Alternate Possibilities," *Journal of Philosophy,* LXVIII (1971), 339-345, *n.* 3 (p. 341).

other than his inability to perform any act other than A_k for the reason described).

Now if Frankfurt has indeed shown that PAP is false, this may be of no great consequence. For it may well be that some trivial modification of PAP is immune to Frankfurt counterexamples and that this modified version of PAP entails that if universal causal determinism and incompatibilism are both true, then all our ascriptions of moral responsibility are false. Frankfurt argues that this is not the case, however, and that what one might call the "correct version" of PAP (that is, the correct principle governing excuse from responsibility in cases in which alternate possibilities for action are absent) cannot be used to show that determinism and moral responsibility are in conflict.[4] I shall not in this paper try to determine whether Frankfurt's proposed principle is true or false, or discuss whether it in fact plays a role in our deliberations about moral responsibility. I shall instead exhibit three principles, which, if they are not "versions" of PAP, are at least principles very similar to PAP, and which *do* play a role in our deliberations about responsibility. I shall argue that these principles are immune to Frankfurt-style counterexamples. (I shall call counterexamples that are directed against principles similar to but distinct from PAP, and which are as strategically similar to Frankfurt counterexamples as is possible, Frankfurt-*style* counterexamples. I shall reserve the term "Frankfurt counterexample" for counterexamples directed just against PAP itself.)

PAP, as Frankfurt formulates it, is a principle about performed acts (things we have done). In Part II, I shall consider a principle about *unperformed* acts (things we have left undone). In Part III, I shall consider two principles about the *consequences* of what we have done (or left undone). In Part IV, I shall argue that if these three principles are true and if a version of incompatibilism appropriate to each is true, then determinism and moral responsibility are in conflict, even given that PAP is false.

[4] The "correct version" of PAP is: "A person is not morally responsible for what he has done if he did it only because he could not have done otherwise" (p. 838).

II

Consider the following principle (the Principle of Possible Action):

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

This principle is intuitively very plausible. But the same might have been said about PAP. Can we show that PPA is false by constructing a counterexample to it that is like Frankfurt's counterexamples to PAP? An adaptation to the case of unperformed acts of Frankfurt's general strategy would, I think, look something like this: an agent is in the process of deciding whether or not to perform a certain act *A*. He decides not to perform *A*, and, owing to this decision, refrains from performing *A*.[5] But, unknown to him, there were various factors that *would have* prevented him from performing (and perhaps even from deciding to perform) *A*. These factors would have come into play if he had shown any tendency towards performing (perhaps even towards deciding to perform) *A*. But since he in fact showed no such tendency, these factors remained mere unactualized dispositions of the objects constituting his environment: they played no role whatever in his deciding not to perform or his failure to perform *A*.

Putative counterexamples to PPA prepared according to this recipe produce, in me at least, no inclination to reject this principle. Let us look at one.

Suppose I look out the window of my house and see a man being robbed and beaten by several powerful-looking assailants. It occurs to me that perhaps I had better call the police. I reach for the telephone and then stop. It crosses my mind that if I do

[5] This schema and the instance of it that follows involve the agent's intentionally refraining from performing a given act. Of course not every case in which we might want to consider holding an agent responsible for failing to perform some act is a case in which the agent intentionally refrains from performing that act: he may never even have considered performing that act. This distinction between two ways of failing to perform a given act is of no importance for our present purposes. The points made in the text would be equally valid if we had chosen to examine a case in which the agent fails even to think of performing the act whose nonperformance we are considering holding him responsible for.

call the police, the robbers might hear of it and wreak their vengeance on me. And, in any case, the police would probably want me to make a statement and perhaps even to go to the police station and identify someone in a lineup or look through endless books of photographs of thugs. And it's after eleven already, and I have to get up early tomorrow. So I decide "not to get involved," return to my chair and put the matter firmly out of my mind. Now suppose also that, quite unknown to me, there has been some sort of disaster at the telephone exchange, and that every telephone in the city is out of order and will be for several hours.

Am I responsible for failing to call the police? Of course not. I couldn't have called them. I may be responsible for failing to *try* to call the police (that much I *could* have done), or for refraining from calling the police, or for having let myself, over the years, become the sort of man who doesn't (try to) call the police under such circumstances. I may be responsible for being selfish and cowardly. But I am simply not responsible for failing to call the police. This "counterexample," therefore, is not a counterexample at all: PPA is unscathed.

It is, of course, proverbially hard to prove a universal negative proposition. Perhaps there are Frankfurt-style counterexamples to PPA. But I don't see how to construct one. I conclude that Frankfurt's style of argument cannot be used to refute PPA.

III

Both PAP and PPA are principles about acts, performed or unperformed. But, in fact, when we make ascriptions of moral responsibility, we do not normally say things like "You are responsible for killing Jones" or "He is responsible for failing to water the marigolds." We are much more likely to say "You are responsible for Jones's death" or "He is responsible for the shocking state the marigolds are in." That is, we normally hold people responsible not for their acts or failures to act (at least explicitly), but for the results or consequences of these acts and failures. What, ontologically speaking, are results or consequences of action and inaction? What sorts of thing are Jones's death and the shocking state the marigolds are in? The general terms "event" and "state of affairs" seem appropriate ones to apply

to these items. But what are events and states of affairs? This question, like all interesting philosophical questions I know of, has no generally accepted answer. Philosophers do not seem even to be able to agree whether events and states of affairs are particulars or universals. In order to avoid taking sides in the debate about this, I shall adopt the following strategy. I shall state a certain principle about excuse from responsibility that seems to me to be a plausible one, provided the events or states of affairs we hold people responsible for are particulars. *And* I shall state a similar principle that seems to me to be plausible, provided the events or states of affairs we hold people responsible for are universals. For each of these principles, I shall argue that it cannot be refuted by Frankfurt-style counterexamples. The first of these principles (which I shall call principles of possible prevention) is:

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

This principle is about events; but if we were to examine a principle, otherwise similar, about "state-of-affairs particulars" (for example, the way secondary education is organized in Switzerland[6]) we could employ arguments that differ from the following arguments only in verbal detail.

What are events if they are particulars? They are items that can be witnessed (at least if they consist in visible changes in visible particulars), remembered, and reported.[7] They are typically denoted by phrases like "the fall of the Alamo," "the death of Caesar," "the death of Caesar in 44 B.C.," and "what Bill saw happen in the garden."[8] How shall we identify and in-

[6] Perhaps it is debatable whether this phrase designates a particular.

[7] I doubt, however, whether they can be anticipated. The objects of anticipation and other "future-directed" attitudes are, I think, universals.

[8] Perhaps the last of these phrases could also be used to name an event-universal. We seem to be using it this way if we say, "What Bill saw happen in the garden happens all too frequently." But, I think, we use it to name a particular when we say, "What Bill saw happen in the garden last night will live in infamy," or "could have been prevented with a little foresight." The phrases "the fall of the Alamo" and "the death of Caesar," however, seem to be suited only for denoting particulars: even if the Alamo had fallen twice, even if Caesar (like Lazarus) had died twice, we could not say, "The fall of the Alamo has happened twice" or "The death of Caesar has happened twice." (This is not

dividuate event-particulars (hereinafter, "events")? Individuating particulars, whether events, tables, or human beings is always a tricky business. (Consider the Ship of Theseus.) As Davidson says,

> Before we enthusiastically embrace an ontology of events we will want to think long and hard about the criteria for individuating them. I am inclined to think we can do as well for events generally as we can for physical objects (which is not very well). . . .[9]

In a paper later than the one this quotation is taken from, Davidson tries to "do as well." He tells us that finding a satisfactory criterion of individuation for events will consist in providing "a satisfactory filling for the blank in:

> If x and y are events, then $x = y$ if and only if _____."[10]

The "filling" he suggests for this blank is (roughly) "x and y have the same causes and effects." The biconditional so obtained, is, I have no doubt, true. But this biconditional will not be "satisfactory" for our purpose, which is the evaluation of PPP1. What we want to be able to do is to tell whether some event that *would* happen if what we earlier called "unactualized dispositions of the objects constituting the agent's environment" were to come into play, is the same as some event (the event responsibility for which we are enquiring about) that actually *has* happened; that is, we want to know how to tell of some given event whether *it*, that very same event, would (nevertheless) have happened if things had been different in certain specified ways. (For when we ask whether an agent could have *prevented* a certain event E by doing, say, X, we shall have to be able to answer the question whether E would *nonetheless* have happened if the agent had done X.)

To see why Davidson's criterion cannot be used to answer our sort of question about event-identity, consider the following formally similar criterion of individuation for persons: "x and

---

due, or not due *solely*, to the presence of the definite article in these phrases, for we can say, "The thing Bill fears most has happened twice.")

[9] From Davidson's contribution to a symposium on events and event-descriptions in *Fact and Existence* ed. by J. Margolis (Oxford, 1969), p. 84.

[10] "The Individuation of Events," in *Essays in Honor of Carl G. Hempel* ed. by N. Rescher (Dordrecht, 1969), p. 225.

y are the same person if and only if x and y have the same blood relatives (including siblings)." This criterion, while *true*, does not help us if we are interested in counterfactual questions about persons. For, obviously, any given man might have had different relatives from those he in fact has (he might have had an additional brother, for example). Davidson's proposed criterion is of no help to us for what is essentially the same reason: any given event might have had different effects from the effects it has in fact had. For example, if an historian writes, "Even if the murder of Caesar had not resulted in a civil war, it would nevertheless have led to widespread bloodshed," he does not convict himself of conceptual confusion. But he is certainly presupposing that the very event we call "the murder of Caesar" might have had different effects.

The above considerations are not offered in criticism of Davidson's criterion, which is, after all, *true*, and may be a very useful criterion to employ (say) when we are asking whether a given brain-event and a given mental event are one event or two. But Davidson's criterion is not the *sort* of criterion we need. We need a criterion that stands to Davidson's criterion as "x and y are the same human being if and only if x and y have the same causal genesis" stands to the above criterion of personal identity. (I use "causal genesis" with deliberate vagueness. A *necessary* condition for x and y having the same causal genesis is "their" having developed from the same sperm and egg.[11] But this is not sufficient, or "identical"—monozygotic—twins would be *numerically* identical.) This criterion can be used to make sense of talk about what some particular person would have been like if things had gone very differently for him.[12] Can we devise a criterion for counterfactual talk about events that is at least no *worse* than our criterion for persons? I would suggest that we simply truncate Davidson's criterion: x is the same event as y if and only if x and y have the same causes. (Note the similarity of this criterion to the causal-genesis criterion of personal identity.) I do not know how to justify my intuition that this criterion is correct, any more

[11] Or so it seems to me. Of course a Cartesian (for example) will have a different view of the matter.

[12] Cf. Saul Kripke, "Naming and Necessity," in *Semantics of Natural Language* ed. by D. Davidson and G. Harman (Dordrecht, 1972), pp. 312-314.

than I know how to justify my belief in the causal-genesis criterion. But, of course, arguments must come to an end somewhere. I can only suggest that since substances (like human beings and tables) should be individuated by their causal origins, and since we are talking about events that, like substances, are particulars, the present proposal is plausible. Moreover, I am aware that this proposed criterion is vague. It is not clear in every case of, say, a story about the events leading up to Caesar's murder, whether it would be correct to say that the murder had "the same causes" in the story that it had in reality. But I think the notion of *same event* is clear just insofar as the notion of *same causes* is clear. And this latter notion is surely not hopelessly unclear: if Cleopatra had poisoned Caesar in 48, then, clearly, there would have happened an event that has not in fact happened, an event that it would have been correct to call "Caesar's death," and which would have had different causes from the event that *is* called "Caesar's death." And, just as clearly, we cannot say of the event we in fact call "Caesar's death," "Suppose *it* had been caused four years earlier by Cleopatra's poisoning Caesar in Alexandria." Moreover, it is hardly to be supposed that we should be able to devise a criterion that will resolve all "puzzle cases," since we are unable to devise such a criterion for people, mountains, or tables.[13]

[13] A theory of event-particulars that is inconsistent with the view presented in this paper is held by R. M. Martin and Jaegwon Kim. (See Martin's contribution to the symposium referred to in *n.* 9, and, for Kim's latest published views on events, "Causation, Nomic Subsumption, and the Concept of Event," *Journal of Philosophy* LXX (1973), 217–236). If we abstract from the particular twists that each of these authors gives to his own account of events, we may say that, on the "Kim-Martin" theory, the class of events is the class of substance-property-time triples. For example Caesar's death is the triple ⟨Caesar, being dead, 15 March 44 B.C.⟩. (Strictly speaking, the term "15 March 44 B.C." in the preceding sentence should be replaced with a term designating the precise instant at which Caesar died.) A "Kim-Martin" event *happens* just in the case that its first term acquires its second term at its third term. However useful Kim-Martin events may be in certain contexts of discussion, I do not think it is correct to think of them as particulars. They are, rather, highly specified universals, just as the property *being the tallest man* is a highly specified (in fact, "definite") universal (cf. *n.* 20). This property, though only one man can have it, is nonetheless such that it *could have* been possessed by someone other than the man who in fact has it. Similarly, any Kim-Martin event that happens *could have* been caused by quite different antecedent events from those that in fact caused it. To suppose that event-particulars have this feature is to violate my

Let us now return to PPP1. Can we devise a Frankfurt-style counterexample to this principle? Let us try.

Gunnar shoots and kills Ridley (intentionally), thereby bringing about Ridley's death, a certain event. 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, since factor F might have caused Ridley's death *by* causing Gunnar to shoot him, perhaps we should say, "if Gunnar had decided not to shoot him"), and (*iii*) is such that Gunnar could not have prevented it from causing Ridley's death except by killing (or by deciding to kill) Ridley himself. So it would seem that Gunnar is responsible for Ridley's death, though he could not have prevented Ridley's death.

It is easy to see that this story is simply inconsistent. What is in fact denoted by "Ridley's death" is not, according to the story, caused by factor F. Therefore, if Gunnar had not shot Ridley, and, as a result, factor F had caused Ridley to die, then there *would have been* an event denoted by "Ridley's death" which had factor F as (one of) its cause(s). But then this event would have been an event other than the event *in fact* denoted by "Ridley's death"; the event in fact denoted by "Ridley's death" would not have happened at all. But if this story is inconsistent it is not a counterexample to PPP1. And I am unable to see how to construct a putative Frankfurt-style counterexample to PPP1 that cannot be shown to be inconsistent by an argument of this sort.

Let us now turn to a principle about universals:

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.[14]

intuitions (at any rate) about particulars. An additional problem: every Kim-Martin event is such that there is some particular moment (its third term) such that the event *must* happen just at that moment if it happens at all. But surely Caesar's death might have happened at least a few moments earlier or later than it in fact did, just as a given man might have been born (or even conceived) at least a few moments earlier or later than he in fact was.

[14] Nothing in PPP1 corresponds to the parenthetical qualification "that state of affairs obtains and" in this principle. So far as I can see, to say of a given event-particular that it "happens" is equivalent to saying that it exists. And, of course, there exist no events that do not exist. Thus there exist no events that do not happen. But states of affairs may exist without obtaining, just as propositions may exist without being true or properties without being instantiated.

The states of affairs "quantified over" in this principle are universals in the way propositions are universals. Just as there are many different ways the concrete particulars that make up our surroundings could be arranged that would be sufficient for the truth of a given proposition, so there are many different ways they could be arranged that would be sufficient for the *obtaining* of a given state of affairs. Consider, for example, the state of affairs that consists in Caesar's being murdered. This state of affairs obtains *because* certain conspirators stabbed Caesar at Rome in 44 B.C., but (since it is a universal), *it*, that very same state of affairs, *might have* obtained because (say) Cleopatra had poisoned him at Alexandria in 48. But this is a bit vague. In order the better to talk about "states of affairs," let us introduce "canonical" names for them. Such names will consist in the result of prefixing "its being the case that" (hereinafter, "C") to "eternal" sentences.[15] Thus a canonical name for the state of affairs referred to above would be "C (Caesar is murdered)." And let us say that the result of flanking the identity-sign with canonical names of states of affairs expresses a truth just in the case that the eternal sentences embedded in these names express equivalent propositions, where propositions are *equivalent* if they are true in just the same possible worlds. (Hereinafter, I shall assume that every proposition is equivalent to and *only* to itself. This assumption could be dispensed with at the cost of complicating the syntax of the sequel.) A state of affairs will be said to *obtain* if the proposition associated with it—that is, the proposition expressed

[15] The choice of eternal sentences as the arguments to which the operator "C" attaches is made largely for the sake of convenience. If we had chosen in addition to eternal sentences, *non*eternal sentences, sentences that can change their truth-values as time passes, for this purpose, then we should have canonical names for states of affairs that can obtain at one time and fail to obtain at other times. If we were to work out a comprehensive and consistent theory of these entities, we should end up with a theory rather like the theory of "states of affairs" R. M. Chisholm presents in "Events and Propositions," *Noûs*, 4 (1970), pp. 15-24. We might, in fact, say that what *we* are calling "states of affairs" are just that subclass of Chisholm's "states of affairs" that he calls *propositions*. If we were to interpret PPP2 as involving quantification over *all* those things Chisholm calls "states of affairs," then (I claim without argument) we could nevertheless defend it against Frankfurt-style counterexamples by arguments essentially the same as those we shall present in the text, but these arguments would be considerably more complicated. For a discussion of the propriety of applying the term "universal" to "states of affairs," see *n.* 20.

by the sentence embedded in any of its canonical names—is true.[16] Thus C(Caesar is murdered), C(Caesar is stabbed), and C(Caesar is poisoned) are three distinct states of affairs, the first two of which obtain and the last of which does not. To *prevent* a state of affairs from obtaining is to prevent its associated proposition from being true (or to *see to it that* or *insure that* that proposition is not true).

Let us now, so armed, return to PPP2. Can we show that PPP2 is false by constructing Frankfurt-style counterexamples to it? What would an attempt at such a counterexample look like? Like this, I think.

Gunnar 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 decided not to shoot him), and (*iii*) is such that Gunnar could not have prevented it from causing Ridley's death except by killing (or by deciding to kill) Ridley himself. So it would seem that Gunnar is responsible for Ridley's being dead though he could not have prevented this state of affairs from obtaining.

This case *seems* to show that PPP2 is false. But in fact it does not. Let us remember that if this case is to be a counterexample to PPP2 and not to some other principle, some principle involving particulars, we must take the words "Ridley's being dead" that occur in it as denoting a universal. What universal? Presumably, C(Ridley dies). But while it is indeed true that Gunnar could not have prevented C(Ridley dies) from obtaining, I do not think it is true that Gunnar is responsible for C(Ridley dies). Why should anyone think he is? Well, Gunnar did something (shooting Ridley) that was *sufficient* for C(Ridley dies). What is more, he performed this act intentionally, knowing that it was sufficient for this state of affairs. This argument, however, is invalid. For consider the state of affairs C(Ridley is mortal). When Gunnar shot Ridley, he performed an act sufficient for (the obtaining of) this state of affairs. But it would be absurd to say that Gunnar

[16] On Chisholm's view (see *n.* 15), the proposition "associated with" a given state of affairs just *is* that state of affairs.

is *responsible* for C(Ridley is mortal). God, or Adam and Eve jointly, or perhaps no one at all, might be held accountable for Ridley's mortality; certainly not his murderer. (Unless, of course, Ridley would have lived forever if he hadn't been murdered; let's assume that is not the case.)

In fact, it is arguable that C(Ridley dies) is the very same state of affairs as C(Ridley is mortal). Given our principle of identity for states of affairs, these "two" states of affairs are one if the two eternal sentences "Ridley dies" and "Ridley is mortal" express the same proposition. And what proposition *could* either of them express but the proposition also expressed by "Ridley does not live forever" and "Ridley dies at some time or other"? So, it should seem, Gunnar is not responsible for C(Ridley dies), and the attempted counterexample to PPP2 fails.

Nor do matters go differently if (somewhat implausibly) we think of "Ridley's being dead" as denoting some more "specific" state of affairs, such as C(Ridley is killed). If Gunnar is indeed responsible for C(Ridley is killed), we shall nevertheless have a counterexample to PPP2 only if Gunnar could not have prevented this state of affairs from obtaining. Let us flesh out "factor F" with some detail to insure that this is the case: suppose there is a third party, Pistol, who would have killed Ridley if Gunnar had not: and suppose Gunnar was able to prevent Pistol's killing Ridley only by killing Ridley himself. By these stipulations, we insure that Gunnar could not have prevented C(Ridley is killed). But do we, in making these stipulations, absolve Gunnar from responsibility for this state of affairs, or is his being responsible for it at least consistent with our stipulations?

I think we absolve him, and that we can show this by an argument of the same sort as the one we used in connection with C(Ridley dies). Let us first note that we cannot show that Gunnar is responsible for C(Ridley is killed) by pointing out that he did something logically or causally sufficient for that state of affairs; for, by the same argument, we could show that he is responsible for C(Ridley is mortal). Now consider the state of affairs—call it "D"—C(either Pistol or Gunnar kills Ridley). Is Gunnar responsible for D? Note that D would have obtained no matter what Gunnar had done, just as C(Ridley is mortal), C(either $2+2=4$ or Gunnar kills Ridley), and C(grass is green or Gunnar kills

Ridley) would have. These latter states of affairs are obviously not ones Gunnar is responsible for. Is there some important difference between them and D in virtue of which Gunnar is responsible for D? There is only one nontrivial difference I can see: There is *no* possible world in which Gunnar is responsible for C(either $2+2 = 4$ or Gunnar kills Ridley); and while there are doubtless possible worlds in which Gunnar is responsible for C(Ridley is mortal) and others in which he is responsible for C(either grass is green or Gunnar kills Ridley), these worlds are exceedingly "remote" from actuality.[17] But some worlds in which Gunnar is responsible for D are much "closer" to actuality than any of these: for example, "close" worlds in which the counterfactual propositions about Pistol that were built into our example are false and Ridley would not have been killed if Gunnar had not shot him. But a miss is as good as a mile; I am arguing only that Gunnar is not *in fact* responsible for D.

Now if Gunnar is not responsible for D, then he is not responsible for C(either Pistol or Gunnar or someone else kills Ridley). And *this* state of affairs and C(Ridley is killed) are one and the same, since the proposition that either Pistol or Gunnar or someone else kills Ridley is equivalent to the proposition that Ridley is killed.[18]

---

[17] Worlds (say) in which Ridley would have lived forever if Gunnar had not shot him, and worlds in which the color of grass is up to Gunnar.

[18] The editors of *The Philosophical Review* have called my attention to the fact that the validity of this argument appears to depend on the doubtful assumption that "Gunnar is responsible for x" is an extensional context. But it need not depend on this assumption. Let us say that in each of the following pairs of sentences the second sentence is a *disjunctive elaboration* of the first.

> All grass is green.
> All grass in London or elsewhere is green.
>
> Ridley is killed.
> Ridley is killed by something or other at some time or other at some place or other.
>
> There is a stack of plates on the table.
> There is a stack of plates on the table that contains twelve plates or else some other number of plates.

Then, I think, a defender of the argument presented in the text need appeal to no principle stronger than: From ⌜S is not responsible for C(p)⌝, derive ⌜S is not responsible for C(q)⌝, provided p is a disjunctive elaboration of q. For

In this example, "factor F" involved a second agent who would have shot Ridley if Gunnar had not. But it would have made no real difference if we had imagined factor F being such that it would have caused Ridley's death by "working through" Gunnar. (See Blumenfeld's counterexample to PAP, quoted in Part I.) Suppose, for example, that Gunnar decides to kill Ridley and does so. Suppose that if he had decided *not* to kill Ridley he would have flushed red (which he couldn't help) and that this red flush together with the prevailing atmospheric conditions would have caused him to decide to kill and, as a result of this decision, to kill, Ridley. Suppose the presence of these atmospheric conditions and the effect on him of their copresence with his flushing red are things he has no choice about. It follows from these suppositions that Gunnar could not have prevented C(Ridley is killed). But we can show by an argument essentially the same as the argument we employed in the "Pistol" case that Gunnar is not responsible for this state of affairs. We proceed by showing first that Gunnar is not responsible for

K C(Ridley is killed by someone who is caused to kill him by factor F [red flush, atmospheric conditions, and so on] or else Ridley is killed by someone who is not caused to kill him by factor F).

This state of affairs plays the role played by C(either Pistol or Gunnar or someone else kills Ridley) in our demonstration that, in the "Pistol" case, Gunnar is not responsible for C(Ridley is killed). We cannot say of K what we said of D, and what we could have said of C(either Pistol or Gunnar or someone else kills Ridley), that it would have obtained no matter what Gunnar had done, for it would not have obtained if Gunnar had not shot Ridley. But we can say of K that it would have obtained no matter what *choices* or *decisions* Gunnar had made, and this seems to me to entail that Gunnar is not responsible for it. (I owe this

example, from "Henry is not responsible for C(There is a stack of plates on the table that contains twelve plates or else some other number of plates)" we derive "Henry is not responsible for C(there is a stack of plates on the table)." This inference seems to me to be plainly valid, even if we suppose Henry to be unable to count beyond three and to be ignorant of the logical principle of Addition.

point to the editors of *The Philosophical Review.*) The remaining step in the demonstration consists simply in observing that the proposition associated with K is equivalent to the proposition that Ridley is killed and, therefore that K and C(Ridley is killed) are one and the same state of affairs, from which fact we infer that Gunnar is not responsible for C(Ridley is killed). (Or, if this inference be thought dubious, we can say that the remaining step consists in observing that the sentence embedded in the displayed name of K is a "disjunctive elaboration" of "Ridley is killed," together with an application of the rule stated in footnote 18.)

If we had chosen to examine instead of C(Ridley is killed) some even more "specific" state of affairs, such as C(Ridley is shot to death at 3:43 PM, 12 January 1949, in Chicago), this would have made no difference to our argument, which in no way depended on the degree of specificity of C(Ridley is killed). An argument of the same sort could be applied to *any* attempt at a Frankfurt-style counterexample to PPP2: the putative counterexample will not be a counterexample *unless* it entails that the agent whose responsibility is in question could not have prevented some given state of affairs; but if the "counterexample" does indeed have this feature, then (I claim) we can always find an argument (sound, I claim), constructed along the lines of the above models, for the conclusion that the agent is not responsible for that state of affairs.

The intuitive plausibility of this conclusion can be shown if we think in terms of the following rather fanciful picture. We are imagining cases in which an agent "gets to" a certain state of affairs by following a particular "causal road," a road intentionally chosen by him in order to "get to" that state of affairs. But, because this state of affairs is a universal, it can be reached by *various* causal roads, some of them differing radically from the road that *is* taken. And, in the cases we imagine, *all* the causal roads that the agent *could* take, all that are *open* to him, lead to this same state of affairs. Perhaps the point of this fanciful talk about "roads" will be clearer if we look at the case of an agent who is unable to prevent a certain state of affairs from obtaining, where this case involves roads in a literal sense. Suppose Ryder's horse, Dobbin, has run away with him. Ryder can't get Dobbin to slow down, but Dobbin will respond to the bridle: whenever

Ryder and Dobbin come to a fork in the road or a crossroads, it is up to Ryder which way they go. Ryder and Dobbin are approaching a certain crossroads, and Ryder recognizes one of the roads leading away from it as a road to Rome. Ryder has conceived a dislike for Romans and so (having nothing better to do) he steers Dobbin onto the road he knows leads to Rome, motivated by the hope that the passage of a runaway horse through the streets of Rome will result in the injury of some of her detested citizens. Unknown to Ryder, however, *all* roads lead to Rome: Dobbin's career would have led him and Ryder to Rome by *some* route no matter what Ryder had done. That is, Ryder could not have prevented C(Ryder passes through Rome on a runaway horse). Is Ryder responsible for this state of affairs? It is obvious that he is not. And it seems obvious that he is not responsible for this state of affairs *just because* he could not have prevented it. I conclude that Frankfurt-style counterexamples cannot be used to show that PPP2 is false.

The universals that PPP2 is "about" are states of affairs; but if we had examined a principle, otherwise similar, about "event-universals" (for example, "its coming to pass that Caesar dies") we could have employed arguments that differed from the above arguments only in verbal detail.

It has been suggested to me[19] that these arguments appear less plausible if one reflects on the fact that essentially similar arguments could be used to show, for example, that Gunnar did not *bring about* C(Ridley is killed) or that Gunnar's pulling the trigger did not *cause* this state of affairs. It is certainly true that if the above arguments are sound, then similar arguments can be used to show that Gunnar did not bring about C(Ridley is killed) and that his bodily movements did not cause this state of affairs to obtain. But these conclusions appear to me to be simply *true*. Let us concentrate on

(1) Gunnar did not bring about C(Ridley is killed).

Why should anyone think (1) is false? It would be clearly invalid to argue that (1) is false since Gunnar did something logically or causally sufficient for C(Ridley is killed), for by the same argu-

[19] By the editors of *The Philosophical Review*.

ment we could establish the falsity of the (true) proposition that Gunnar did not bring about C(Ridley is mortal). Or consider the case of Ryder and Dobbin. In turning down a certain road, Ryder did something causally sufficient for passing through Rome on a runaway horse, but would anyone want to say that Ryder brought about the (for him inevitable) state of affairs C(Ryder passes through Rome on a runaway horse)?

The states of affairs we have been considering are *universals*. There are *many* ways the concrete particulars that make up our surroundings could be arranged that would be sufficient for their obtaining. What Gunnar and Ryder can bring about is *which* of these possible arrangements of particulars (which murderer, which road) the universals will be "realized in"; that *some arrangement or other* of the particulars will realize these universals is something totally outside their control; it is not something they bring about. Here is an analogy involving another sort of universal, properties. Chisel is a sculptor and sculpts the heaviest statue that ever was or will be, *The Dying Whale*. Thus Chisel brings it about that a certain particular, *The Dying Whale*, exemplifies the property of being the heaviest statue.[20] But he does not bring

[20] Perhaps some philosophers would be disinclined to call the property of being the heaviest statue there ever was or will be a *universal*, on the ground that a universal must be "sharable," must be capable of being exemplified by more than one object. And, for similar reasons, it might be held that what I have called "states of affairs" are not true universals, since each of them either obtains or fails to obtain without further qualification, whereas a state of affairs that was truly a universal should be capable (say) of obtaining in 1943 but not in 1956 (cf. *n.* 15), or of obtaining in both Britain and the United States but not in France. Well, let us say that our "states of affairs" and properties like being the heaviest statue are, if not "true" universals, at least *cross-world universals*. A property or other abstract object is a cross-world universal if there are worlds W_1 and W_2 such that x falls under it in W_1 and y falls under it in W_2 and $x \neq y$. (I use the words "fall under" with deliberate vagueness; what "falls under" a property is whatever has it; what "falls under" a state of affairs is whatever arrangement of particulars realizes it.) If this usage is an extension of traditional philosophical usage, it is a very natural one; I call, e.g., C(Gunnar kills Ridley) a "universal" because it is not "tied to" any given arrangement of particulars. I do not pretend that these remarks are very precise. Certainly the notion of an "arrangement of particulars" could do with some clarification. For example, it is not clear what should be said about states of affairs that, unlike those discussed above, involve only a single particular. (Let us say that a state of affairs *involves* a particular if that particular is such that its existence is entailed by the obtaining of that state of affairs.) Consider, for example, C(there is such a building as the Taj Mahal). Are there many "arrangements

it about that this property is exemplified, since, no matter what he had done, this property would "automatically" have been exemplified by something or other: he causes something to exemplify this property, but he does not cause this property to be exemplified.

In affirming (1), I do not mean to affirm the falsehood

(2) Gunnar did not bring about Ridley's death,

where "Ridley's death" denotes an event-*particular* (individuated from other particulars in virtue of having different causal antecedents), one that is also perhaps denoted by "Ridley's death on Thursday," "the only death Gunnar ever caused," and so on. Anyone who feels inclined to reject (1) should make sure that this inclination does not arise from a failure to distinguish between (1) and (2). To revert to the sculpture example, (1) and (2) stand to each other roughly as

Chisel did not cause the property of being the heaviest statue to be exemplified,

and

Chisel did not cause (the particular thing that is) the heaviest statue to exist,

---

of particulars" in which this state of affairs could be realized? Tentatively, I should say Yes. I should think that "the arrangement of particulars that realizes a given state of affairs" should in general be taken to be an arrangement of a broader class of particulars than those it "involves." For example, C(there are humans) does not in the strict sense defined above *involve* you or me (in fact, *no* contingent being is such that this state of affairs involves it), but you and I are, in a very intuitive sense, among those particulars the arrangement of which realizes it. Similarly, though no block of marble is such that C(there is such a building as the Taj Mahal) involves it—at least on the assumption that mereological essentialism is false—many blocks of marble would seem to be among those particulars the arrangement of which realizes it. Or even if we do not consider *parts* of the Taj Mahal, we must admit that the state of affairs we are considering would obtain if the Taj Mahal were differently placed or differently oriented; and it seems intuitively correct to say that if the place or orientation of the Taj Mahal were different from what it in fact is, then C(there is such a building as the Taj Mahal) would be realized in a different arrangement of particulars.

stand to each other. The former is, as I argued above, true, and the latter false.[21]

So, it would seem, we are unable to devise a Frankfurt-style counterexample either to PPP1 or to PPP2. If our attempts at counterexamples looked initially plausible, this, I think, was due to a confusion. When we hear the Gunnar-Ridley story, it *seems* correct to say that it follows from the story that Gunnar is responsible for Ridley's death *and* that Gunnar could not have prevented Ridley's death. But "Ridley's death" is ambiguous. If we are using this phrase to denote a universal, then we may say that Gunnar could not have prevented Ridley's death, but not that he was responsible for Ridley's death. If we are using this phrase to denote a particular, then we may say that Gunnar was responsible for Ridley's death, but not that he could not have prevented it.

This result might lead us to wonder whether Frankfurt's counterexamples to PAP rest on a similar confusion. Suppose we were to split PAP into two principles, one about "act-particulars" (event-particulars that are voluntary movements of human bodies) and one about "act-universals" (that is, things that could be done by distinct agents, such as murder, prayer, or killing Jones at noon on Christmas Day, 1953): should we then see that Frankfurt's alleged counterexamples to PAP depend for their plausibility on treating one and the same act as a particular at one point in the argument, and a universal at another?

I do not think that Frankfurt is guilty of any such confusion. The "acts" that figure in his counterexamples seem to me to be treated consistently as universals. If this is the case, it raises two questions. Let us split PAP into two principles as was suggested in the preceding paragraph: PAP1 (about particulars) and PAP2 (about universals). The first question: If indeed Frankfurt's "acts" are universals, he is arguing against PAP2; can his argument be met by considerations like those we raised in defense of PPP2? The answer seems to me to be No, but I am not at all sure about

[21] I do not mean to give the impression that one never brings about any state of affairs. For example, (granting the correctness of the Warren Commission Report), Lee Harvey Oswald brought about Ç(Kennedy dies on 22 November 1963). But it is *not* true that Oswald brought about C(Kennedy dies). That state of affairs was brought about by God or by Adam and Eve or by no one at all. Moreover, it *is* true that Oswald brought about the event-particular, Kennedy's death.

this. The considerations raised in defense of PPP2 depended on our having at our disposal a fairly precise notion of "state-of-affairs universal," and I am not at present able to devise an equally precise notion of "act-universal" that I find satisfactory.[22] The second question: what about PAP1? I do not find this question interesting, since I do not think that "event-particulars that are voluntary movements of human bodies" are what we hold people responsible for. I shall not, however, defend this view here. An adequate defense of it would be fairly complex, and I do not think my reasons for thinking what I do on this matter are worth developing merely to establish a negative conclusion.

---

[22] An adequate construction of such a notion would require the introduction of a canonical language for act-universals. I am unable to devise a language for this purpose that comes close to satisfying me. Even without having such a language at my disposal, however, I think I see a serious obstacle to any attempt to refute Frankfurt's arguments against PAP2 by raising considerations like those used to defend PPP2 in the text. Let us suppose that "the act of killing Ridley" denotes a certain act-universal, an act such that *it*, that very act, could be the act of any among a number of agents and be performed under a great variety of conditions. Consider the following Frankfurt counterexample to PAP2: Gunnar performs the act of killing Ridley; moreover, if he had decided not to perform it, some third party, Cosser, would have caused him to perform it. If we were to try to refute this counterexample by arguments parallel to those we used in defense of PPP2, we should have to find an act-denoting phrase that stands to "the act of killing Ridley" roughly as "C(either Pistol or Gunnar or someone else kills Ridley)" stands to "C(Ridley is killed)." I am not sure what such a phrase would look like, but I think something like this:

> The act of killing Ridley, either without having been caused to kill Ridley by anyone, or as a result of having been caused to kill Ridley by Cosser or someone else.

I am very doubtful whether this phrase makes any sense. To take a simpler case, given that there is such an act as eating forbidden fruit, an act one might perform as a result of one's having been given bad advice, is there such an act as the act of eating forbidden fruit as a result of having been given bad advice? I find the notion of such an act difficult to grasp. But if no coherent act-universal-name can be found to play the formal role played by "C(either Pistol or Gunnar or someone else kills Ridley)" in our defense of PPP2, then no parallel argument in defense of PAP2 can be constructed.

These considerations, of course, do not show that Frankfurt's attack on PAP is successful. They do, however, raise serious doubts about the possibility of defending PAP against this attack by constructing an argument formally parallel to our argument in defense of PPP2.

IV

We have shown that three principles relating ability and responsibility cannot be refuted by Frankfurt-style counter-examples:

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 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.

Now consider three versions of incompatibilism:

> If determinism is true, then if a given person failed to perform a given act, that person could not have performed that act.
>
> If determinism is true, then no event is such that anyone could have prevented it.
>
> If determinism is true, then if a given state of affairs obtains, then no one could have prevented that state of affairs from obtaining.[23]

Obviously, if these three theses are true, then (since PPA, PPP1, and PPP2 are true) it follows that determinism entails that no one has ever been or could ever be responsible for any event, state of affairs, or unperformed act. Moreover if the following schema

R If S is responsible for Φing, then there is some event or state of affairs for which S is responsible,

[23] I think I am justified in calling these three theses "versions" of a single doctrine, since, *if* there were a good argument for any of them, then, I should think, it could be easily modified to yield a good argument for either of the others. I have presented arguments for what is essentially the first of these three versions of incompatibilism in "A Formal Approach to the Problem of Free Will and Determinism," *Theoria* XL (1974) Part 1, pp. 9-22, and "The Incompatibility of Free Will and Determinism," *Philosophical Studies* 27 (1975) pp. 185-199.

(here "Φing" is to be replaced by any grammatically appropriate action phrase) is valid, then determinism is (assuming incompatibilism) incompatible not only with our being responsible for the consequences of our acts but for our acts themselves. And this schema is extremely plausible. I cannot myself conceive of a case in which an agent is responsible for having performed some act but is responsible for *none* of the results or consequences (either universal or particular) of this act.[24]

Thus, if all three versions of incompatibilism are true, and if determinism is true, then there is simply no such thing as moral responsibility. There is such a thing as moral responsibility only if someone is responsible for something he has done, or for something he has left undone, or for the results or consequences of what he has done or left undone. And the principles for which I have argued (PPA, PPP1, PPP2, and the validity of schema R) entail that if incompatibilism is true, then determinism is incompatible with anyone's being responsible for anything whatever.

Therefore, even if PAP is false,[25] and even if Frankfurt's "correct version" of PAP (see footnote 4) cannot be used to show that determinism and moral responsibility are incompatible, it is *nonetheless* true that unless free will and determinism are compatible, determinism and moral responsibility are incompatible. Thus, Frankfurt's arguments do not, even if they are sound, rob the compatibilist-incompatibilist debate of its central place in the old controversy about determinism and moral responsibility.[26]

Syracuse University

---

[24] An obvious argument for the validity of R is this: If someone Φs and is responsible for so acting, then, whatever other events or states of affairs he may be responsible for, he is at least responsible for its being the case that he Φs. But this argument is unsound. Consider the case (p. 215 above) involving the counterfactual propensities of atmospheric conditions to cause Gunnar to decide to kill, and to kill, Ridley. I argued that in that case Gunnar is not responsible for C(Ridley is killed). A similar argument could be used to show that in that case Gunnar is not responsible for C(Gunnar kills Ridley). But it does not follow that Gunnar is not responsible for killing Ridley. For Gunnar might have freely decided to kill Ridley and have killed him as a result of this free decision (and thus be responsible for killing Ridley); nevertheless, *if* he had (freely) decided *not* to kill Ridley, external factors outside his control

would *then* have "come into play" and caused him (unfreely, of course) to kill Ridley. Therefore, while Gunnar is responsible for killing Ridley, he is not responsible for C(Gunnar kills Ridley freely or Gunnar kills Ridley unfreely) and hence is not responsible for C(Gunnar kills Ridley). Thus our "obvious" argument for the validity of R is fallacious.

Nonetheless, R seems to me to be valid. Certainly the case we have just considered is not a counterexample to its validity. For, in this case, while Gunnar is not responsible for C(Gunnar kills Ridley), he *is* responsible for C(Gunnar kills Ridley without having been caused to do so by atmospheric conditions). Moreover, he is responsible for the event-particular, Ridley's death.

[25] Of course, if the above arguments are correct, and if determinism and incompatibilism are both true, then PAP *is* true: it is vacuously true because no one, in that case, is responsible for anything he does. Frankfurt, of course, does not mean to deny that PAP might be, as a matter of contingent fact, vacuously true.

[26] I should like to thank the editors of *The Philosophical Review* for their careful comments on earlier versions of this paper, which have led to many improvements. I am especially grateful to them for pointing out to me that an argument I employed was invalid.