DAVID P. HUNT

MORAL RESPONSIBILITY AND UNAVOIDABLE ACTION

(Received 28 July 1997)

ABSTRACT. The "principle of alternate possibilities" (PAP), making the ability to do otherwise a necessary condition for moral responsibility, is supposed by Harry Frankfurt, John Fischer, and others to succumb to a peculiar kind of counterexample. The paper reviews the main problems with the counterexample that have surfaced over the years, and shows how most can be addressed within the terms of the current debate. But one problem seems ineliminable: because Frankfurt's example relies on a "counterfactual intervener" to preclude alternatives to the person's action, it is not possible for it to preclude *all* alternatives (intervention that is contingent upon a trigger cannot bring it about that the trigger never occurred). This makes it possible for the determined PAPist to maintain that some pre-intervention deviation is always available to ground moral responsibility.

In reply, the critic of PAP can examine all the candidate deviations and argue their irrelevance to moral responsibility (a daunting prospect); or the critic can dispense with counterfactual intervention altogether. The paper pursues the second of these strategies, developing three examples of noncounterfactual intervention in which (i) the agent has no alternatives (and *a fortiori* no morally relevant alternatives), yet (ii) there is just as much reason to think that the agent is morally responsible as there was in Frankfurt's original example. The new counterexamples do suffer from one liability, but this is insufficient in the end to repair PAP's conceptual connection between moral responsibility and alternate possibilities.

Conventional wisdom links moral responsibility with the power to control one's actions: absent any possibility of avoiding an action, an agent cannot be held accountable for proceeding with the action. Harry Frankfurt, in an influential paper,[1] called this bit of wisdom the "principle of alternate possibilities," or 'PAP', and formulated it as follows:

PAP: A person is morally responsible for what he has done only if he could have done otherwise.

Philosophical Studies **97:** 195–227, 2000.

Frankfurt noted that PAP "has generally seemed so overwhelmingly plausible that some philosophers have even characterized it as an *a priori* truth";[2] proof of its privileged standing lies in the fact that it is the common property of compatibilists and incompatibilists alike, who differ only in how they understand 'could'. Despite the secure position it appears to occupy, however, a number of philosophers, including Robert Nozick, John Martin Fischer, and Frankfurt himself, have argued that PAP is in fact false.[3] Their contention that an agent can be morally responsible even for actions that are unavoidable represents one of the most important challenges to conventional wisdom to come out of recent work on free will.

The centerpiece of Frankfurt's attack on PAP is a celebrated counterexample. David Blumenfeld, in the earliest published response to Frankfurt, claimed that "[t]he argument against the principle rests, essentially, on a counterexample,"[4] and the attention this counterexample has attracted in the subsequent literature is ample testimony to the pervasiveness of this judgment. It is not surprising, given the intense scrutiny it has received over the years, that a number of potential problems with the counterexample have been identified. The import of these problems has of course been disputed by defenders and critics of Frankfurt's argument. But because the counterexample has been identified so closely with the case against PAP, both sides have tended to see the adjudication of these problems as critical to the case itself and not just to the counterexample's usefulness. In this respect the exclusive focus on the counterexample has been unfortunate, since it can lead one to forget that the basic insight the counterexample is supposed to serve is independent of Frankfurt's original example and any infelicities it may involve.

In what follows, I argue that the case against PAP has been unnecessarily handicapped by Frankfurt's choice of counterexample, inasmuch as the main objections to it take advantage of inessential features of the counterexample; and I show that when these inessential features are purged, the standard objections collapse. To this end, I begin by presenting Frankfurt's counterexample and the argument it is intended to serve. I then consider the principal objections that have been raised against the argument, showing to what extent these can be met while retaining the basic features of Frankfurt's original counterexample and to what extent those features must be modified

if the objections are to be finessed. I conclude by developing some improved counterexamples and reviewing the status of PAP in light of these counterexamples.

FRANKFURT'S ARGUMENT AGAINST PAP

Suppose that an agent S performs an action A at a time t, and that S cannot avoid performing A at t. These are two distinct (though obviously related) facts. Frankfurt's argument against PAP is rooted in the idea that, because these two facts are distinct, we can distinguish (at least verbally) between the circumstances responsible for the first fact and the circumstances responsible for the second fact.

In the typical case of unavoidable action the two sets of circumstances coincide, and the distinction is merely verbal. Here "the same circumstances both bring it about that a person does something and make it impossible for him to avoid doing it," Frankfurt explained, citing as examples "situations in which a person is coerced into doing something, or in which he is impelled to act by a hypnotic suggestion, or in which some inner compulsion drives him to do what he does."[5] PAP entails that the person is not morally responsible in such cases, and here it clearly delivers the right result. But we can also conceive of the two sets of circumstances diverging from each other. "A person may do something in circumstances that leave him no alternative to doing it," Frankfurt suggested, "without these circumstances actually moving him or leading him to do it – without them playing any role, indeed, in bringing it about that he does what he does."[6] It is in unusual situations such as this, where the circumstances responsible for the person's action and the circumstances responsible for the action's inevitability go their separate ways, that Frankfurt thought PAP might be vulnerable. In such a case the person might find his action unavoidable in virtue of one set of conditions while remaining morally responsible for the action in virtue of a distinct set of conditions. The conditions making for unavoidability, since they contribute nothing toward the actual performance of the action, can do nothing to undermine the person's moral responsibility for the action (supposing such responsibility to be otherwise warranted by the facts of the case).

There are clearly the makings here of a purely theoretical argument against PAP, one which argues for the moral irrelevance of an action's unavoidability when the latter makes no actual contribution toward the action's performance. Instead of arguing directly in this manner, however, Frankfurt chose to develop a counterexample. Here is a version of Frankfurt's counterexample. Jones murders Smith, and does so under conditions which would normally entail Jones's moral responsibility for the murder. (Take your favorite theory of moral responsibility and imagine that its clauses are all satisfied by Jones's action, at least insofar as their satisfaction is consistent with the other details of the counterexample.) Conditions are not entirely normal, however. Lurking in the background is a third party, Black, who wishes Jones to murder Smith, and who possesses a mechanism capable of monitoring and controlling a person's thoughts. Black could use the mechanism to force Jones into murdering Smith, but he decides against taking such a direct approach. Black, as Frankfurt puts it, "prefers to avoid showing his hand unnecessarily";[7] what's more, he fully expects that intervention will *be* necessary, since Jones has ample reasons of his own for wishing to see Smith dead. But Black is also unwilling to be disappointed in this expectation. He therefore adopts a plan designed to guarantee the murder while minimizing his own involvement: he programs the mechanism to monitor Jones's thoughts for evidence of his intentions with respect to murdering Smith, and to manipulate those thoughts to ensure the murder of Smith if (and only if) it appears that Jones is not going to acquire the requisite intention in any other way. As it happens, the mechanism does not have to intervene in the course of events, because Jones goes ahead and murders Smith on his own.

Frankfurt thought that two judgments are pretty obviously warranted in such a case. First, Jones is morally responsible for killing Smith: the other aspects of the situation were posited to be ideal for moral responsibility, while the mechanism did not end up contributing in any way to Jones's decision to kill Smith, which would have occurred just as it did even if the mechanism had not existed. (Jones could hardly deflect responsibility from himself by claiming, "The mechanism made me do it.") Second, Jones was unable to refrain from killing Smith: given the mechanism,

no alternative course of action was available to Jones, though he was completely unaware of this fact. The case therefore seemed to Frankfurt to constitute a decisive counterexample to PAP.

Call this the 'Black-Smith-Jones' counterexample, or 'BSJ' for short. Frankfurt's argument, and BSJ's role in it, appears to be of the following sort:

(1) If there is a case in which (i) S is morally responsible for A-ing at t and (ii) S cannot do otherwise than A at t, then PAP is false.

(2) BSJ is a case in which (i) S is morally responsible for A-ing at t and (ii) S cannot do otherwise than A at t.

(3) There *is* a case in which (i) S is morally responsible for A-ing at t and (ii) S cannot do otherwise than A at t.

(4) Therefore, PAP is false.

The argument is clearly valid. If it is nevertheless unsound, the problem must be located in premise (2), since (1) is analytic of PAP and (3) follows from (2) by existential generalization. It is not surprising, then, that BSJ has been the focus of the argument's critics, who have tried to show that (contrary to appearances) the conditions obtaining in BSJ are such that either S is *not* morally responsible or S *can* do otherwise. But premise (1) may also be contestable – not on grounds of truth, to be sure, but on grounds of relevance. In particular, critics might question whether PAP is an apt expression of the "alternate possibilities" required by moral responsibility. In the remainder of this section I make some brief remarks on the logic of these two challenges, beginning with concerns over premise (1).

There is an ambiguity in PAP, which some critics have fastened onto, over *what* S is supposed to be able to do otherwise, and *when* he is supposed to be able to do it. It may not be clear, for example, whether the intended requirement is

PAP1: S is morally responsible for A-ing at t only if (at t) he could have done other than A at t

or

PAP2: S is morally responsible for A-ing at t only if (at some time u) he could have (so acted that he would have) done other than A at t.

Likewise, PAP could be read as affirming that

PAP3: S is morally responsible for A-ing at t only if he could have *done something else instead of A-ing at t*

or simply asserting that

PAP4: S is morally responsible for A-ing at t only if he could have *refrained from A-ing at t.*

PAP1 is a special case of PAP2, and PAP3 a special case of PAP4. As there is little reason to think that the special case is a necessary condition for moral responsibility (PAP1 in particular being a non-starter: it is often too late at t to prevent an action initiated earlier), we can resolve these ambiguities by combining PAP2 and PAP4 into

PAP*: S is morally responsible for A-ing at t only if (at some time u) he could have (so acted that he would have) refrained from A-ing at t.

Whenever I use 'PAP' in the narrow and strict sense in this paper, it is PAP* that I should be understood to have in mind.

In addition to this proper sense of 'PAP', which seems the most plausible expansion of the formula Frankfurt actually gives, we might ask whether another principle isn't also (supposed to be) at stake here, namely,

PAP+: S is morally responsible for A-ing at t only if (there is something) he could have done otherwise (and it is at least in part in virtue of what he could have done otherwise that he is morally responsible for A-ing at t).

This is a broader principle which includes PAP itself as a special case, namely, the case in which what the person could have done otherwise (and in virtue of which he is morally responsible for A-ing at t) is the very act of A-ing at t. The last parenthetical clause in PAP+ is required by the existence of morally irrelevant alternative possibilities. If Jones is responsible for murdering Smith (at least in

part) in virtue of a power to do otherwise, it cannot be in virtue of his having had the power to eat sausage instead of bacon for breakfast, or to refrain from blinking his eyes as he raised the murder weapon. The existence of a morally irrelevant alternative should not even be a necessary condition for moral responsibility.[8]

Is this broader principle the real target of Frankfurt's argument? PAP+ certainly represents a significant departure from the simple idea behind PAP: that moral responsibility for A-ing is tied to the power to refrain *from A-ing*. This is one reason not to conflate PAP+ with PAP. Another is that BSJ does not speak to PAP+ in anything like the transparent way it speaks to PAP; so insofar as one can infer Frankfurt's target from the weapon he wields against it, there is some further reason for doubting that this target is PAP+. Nevertheless, Frankfurt's argument would lose much of its interest if restricted to the narrower PAP. Where A is an overt action like Jones's shooting Smith, A is the end-product of a series of events originating in an act of will. The latter is the primary locus of both moral responsibility and control. Nothing so complicated as BSJ is needed to show that the control I have over my own will (when morally responsible) weakens with each step in the casual chain my will originates. I cannot be held morally responsible for whether my body works properly or whether the external world is cooperative (unless of course I contributed in some way to their intractability). External events may conspire to suppress the intended outcome(s) of my agency without my responsibility being thereby annulled. It is presumably Frankfurt's recognition that the real game is the inner one that leads him to characterize Black's mechanism as one that hijacks Jones's mental processes – as opposed, say, to one that manipulates only his gross bodily movements, leaving him the helpless spectator of outward behavior over which he has no control. There is some reason, then, to think that Frankfurt's argument is directed against alternate possibilities *in general* as a requirement of moral responsibility, and not just the alternatives covered under a narrow (PAP*-like) reading of PAP.

Rather than deciding prematurely the intended scope of Frankfurt's attack, I prefer to regard principles which require alternate possibilities for moral responsibility, but formulate the requisite alternative differently than PAP, as potential targets for Frank-

furt's argument. Some sort of alternate possibilities requirement has been axiomatic in conventional thinking about free agency and moral responsibility, and it is this requirement that Frankfurt means to be challenging. If PAP is not a perspicuous rendering of that requirement, then PAP can at least stand proxy for the requirement (whatever it might be). In contrast, strictures on moral responsibility which, while assigning a role to alternatives or the lack thereof, do not require that there *be* alternate possibilities, must certainly be treated as an abandonment of PAP. These include

> A person is morally responsible for what he has done only if he did not do it *because* he could not have done otherwise.[9]

and

> A person is morally responsible for what he has done only if he did not do it *only* because he could not have done otherwise.[10]

There is no need to consider these substitutes for PAP in what follows, since they do not require that there *be* any alternate possibilities.

Suppose that Frankfurt's PAP fails to capture the most plausible version of the alternate possibilities requirement. What implications would this have for Frankfurt's argument? None at all, so long as the relevance of the argument's conclusion is unimportant. But if relevance is a consideration and premise (1) is modified accordingly, the resulting argument would likely be invalid (the new premise (1) wouldn't connect with the old premise (2) in the right way to produce a valid *modus ponens* inference). This is an insufficient basis for dismissing the concerns raised by Frankfurt's argument, however, since one could presumably construct a new version of (2) to match the new version of (1), and there may be just as much reason to accept the new (2) as there was to accept the old (2). So it looks like the endgame will come down to an assessment of (2) in any case, whether in its unmodified form or under some modification required by a reassessment of (1).

Let us turn then to premise (2). BSJ is supposed to instance the general scenario Frankfurt formulated at the beginning of his

paper, consideration of which (he predicted) would "cast doubt ... on the relevance to questions of moral responsibility of the fact that a person who has done something could not have done otherwise." The general scenario may be characterized more carefully as follows. Let '$\overline{\text{PAP}}$' designate all the necessary conditions for moral responsibility other than PAP.[11] (PAP plus $\overline{\text{PAP}}$ are then *sufficient* for moral responsibility.) Now let A be an action that an agent S performs at a time t, where S's A-ing at t satisfies $\overline{\text{PAP}}$; and let C_α comprise all the conditions which actually contributed toward S's A-ing at t. Suppose further that there is a set of conditions C_N, disjoint from C_α, given which S cannot do otherwise than A at t. The foregoing specifications, including most prominently the presence of a C_N which is disjoint from C_α, define what I shall call a 'Frankfurt scenario'.[12]

Though the notion of a Frankfurt scenario plays no direct role in Frankfurt's finished argument against PAP, it enters indirectly at two points. First, Frankfurt's sketch of the scenario at the beginning of his paper raises initial doubts about PAP's invulnerability, and suggests a direction in which counterexamples to PAP might be found. The Frankfurt scenario, in effect, provides a template for the construction of BSJ; the utility of the template is then confirmed when BSJ turns out to constitute a counterinstance to PAP. Second, once BSJ is available for step (2) of the argument, the notion of a Frankfurt scenario renders salient those features of BSJ in virtue of which (2) is true: it is true *because* the conditions making for unavoidability were not part of the actual sequence. The first point belongs to the "logic of discovery"; but the second is arguably part of the "logic of justification," and points toward a sub-argument in support of (2):

(2a) Any case constituting a Frankfurt scenario is a case in which (i) S is morally responsible for A-ing at t and (ii) S cannot do otherwise than A at t

(2b) BSJ is a case constituting a Frankfurt scenario

(2) BSJ is a case in which (i) S is morally responsible for A-ing at t and (ii) S cannot do otherwise than A at t

One virtue of regimenting the argument this way is that it distinguishes (2b) the *factual* judgment that certain circumstances obtain

from (2a) the *moral* judgment that is called for in those circumstances. This ensures that objections to premise (2) will fall into one or the other category as well.

Critics who attack the truth of (2), rather than the relevance of (1), have focused overwhelmingly on (2b) as the problematic move, while leaving (2a) alone. Given a Frankfurt scenario, with its unavoidable outcome, is S nevertheless morally responsible for A-ing at t, or does the absence of alternate possibilities preclude moral responsibility, even though C_α and C_N diverge? Answering this question means adopting a substantive moral position, and on this substantive point Frankfurt's critics appear to take his side: the pivotal moral intuition in Frankfurt's argument is shared (or at least unchallenged) by the argument's critics. As this will be a point of some importance later on, it is worth flagging this intuition for future reference. Let us call it the 'Master Intuition', and formulate it as follows:

(MI) Were S to A at t as part of a Frankfurt scenario – i.e., in circumstances such that S's A-ing at t satisfies $\overline{\text{PAP}}$, there are conditions C_N making S's A-ing at t unavoidable, and $C_N \neq C_\alpha$ – S would be morally responsible for A-ing at t.

This intuition can be rejected, to be sure – but a *simple* rejection, without any argument to undermine the intuition, does nothing to refute Frankfurt's case against PAP. Clearly anyone who fails to share this intuition will fail to accept Frankfurt's argument, but it's not clear why this should count as a problem with the argument: any argument rests on assumptions, and Frankfurt's rests on (MI). For Frankfurt's argument to raise serious worries regarding PAP, it is enough that (MI) be prima facie plausible, and that its prima facie plausibility not be undercut by any secunda facie counterarguments. This standard it appears to meet. In any case, Frankfurt's critics have been content to leave (MI) unchallenged.

What critics dispute instead is the argument's factual premise asserting that BSJ constitutes a genuine Frankfurt scenario. David Widerker is typical in this regard. After quoting Frankfurt's claim that "there may be circumstances that make it impossible for a person to avoid performing some action, without those circumstances in any way bringing it about that he performs that action,"[13]

Widerker suggests that "the success of Frankfurt's case against PAP depends crucially upon his ability to convince us of the plausibility of [this claim]."[14] The problem with Frankfurt's argument, as Widerker develops it, is not that Frankfurt scenarios turn out to satisfy PAP after all (vitiating the argument's moral premise), but that in constructing BSJ "Frankfurt has failed to give us an example"[15] of a Frankfurt scenario (the argument's factual premise). In sum, critics appear to grant that Frankfurt scenarios (supposing that there were any) would be counterexamples to PAP; but they deny that BSJ is a genuine Frankfurt scenario. If the critics are right in this judgment, BSJ must fall short, either because it fails to exclude all alternatives or because it fails to preserve those features of the Frankfurt scenario which allow the latter to maintain moral responsibility in the face of unavoidability. In the first section to follow I look at the latter criticism, and in the second at the former.

One final point. Anyone looking to (2b) for the flaw in Frankfurt's argument against PAP must hold, not only that BSJ fails to satisfy all the specifications for a Frankfurt scenario, but that nothing does (could) satisfy them. Otherwise a new case can be constructed in support of (3). (This would simply be a continuation of the dialectic in Frankfurt's original article, which moved from less to more adequate counterexamples to PAP, culminating in BSJ. Rejecting (2b) might simply invite a further modification in BSJ rather than a reaffirmation of PAP.) By and large, critics do little to argue on behalf of this stronger conclusion. Perhaps the idea is that, since PAP is strongly intuitive, the burden of proof falls on Frankfurt and his supporters. Since Frankfurt has put BSJ forward as the prize counterexample to PAP, it is enough to defeat this counterexample for the presumption to revert in favor of PAP. The critics may well be right in this assessment of the dialectical situation. In any case, it invites the skeptic to propose alternative counterexamples in the event that (2b) is refuted. This will be my procedure in the next three sections of the paper. Instead of counterarguing that BSJ *is* a genuine Frankfurt scenario after all, I shall concede BSJ's inadequacies (at least for the sake of argument) and develop alternative counterexamples by which (3) can be reached without reliance on (2b).

THE FIRST OBJECTION AND AN IMPROVED COUNTEREXAMPLE

The first objection I want to consider is that BSJ, with the mechanism operating as described, covertly presupposes causal determinism. Why might someone think this? The reason has to do with the fact that in BSJ intervention occurs when (and only when) necessary to ensure S's A-ing at t. Now intervention is necessary whenever S would otherwise refrain from A-ing at t. The mechanism as described therefore requires that there be a set of triggers {T} such that – absent intervention by the mechanism – were some member of {T} to obtain, S would refrain from A-ing at t, and were no member of {T} to obtain, S would A at t. But this appears to make S's action causally dependent on prior conditions (the occurrence or nonoccurrence of the members of {T}). Once this presupposition of causal determinism is brought to light, the libertarian (at least) will want to deny that BSJ is after all a case in which the "agent" is morally responsible – so (2) is false. Moreover, it is false because (2b) is false. A Frankfurt scenario must satisfy $\overline{\text{PAP}}$ – indeed, it is only to the extent that it satisfies $\overline{\text{PAP}}$ that such a scenario is compatible with moral responsibility. But for a libertarian, $\overline{\text{PAP}}$ will include causal indeterminism. Since BSJ excludes causal indeterminism, it cannot constitute a Frankfurt scenario.[16]

One response, following Blumenfeld, is that the members of {T} might be *signs* of S's future A-ing at t rather than causal determinants of it. Another is that the subjunctive conditionals linking the members of {T} with S's refraining from A-ing at t (and the nonobtaining of {T} with S's A-ing at t) might be Molinist "counterfactuals of freedom" rather than causal regularities. Both responses have been attacked and defended in a recent exchange between David Widerker and John Fischer.[17]

Rather than joining that exchange, I would like to suggest that the whole discussion can be avoided inasmuch as it rests on a vulnerability that is not essential to the counterexample's effectiveness. To play its assigned role in refuting PAP, it is necessary only that the mechanism intervene whenever Jones would otherwise desist from murder, and *not* intervene given the actual course of affairs. It is not also necessary that the mechanism intervene *only* when Jones would otherwise desist. The latter requirement reflects Frankfurt's ascribing to Black a wish "to avoid showing his hand unneces-

sarily." Why did Frankfurt include this detail in his description of BSJ? Its inclusion does serve to explain why Black's mechanism does not actually intervene (given that its intervention is actually unnecessary); but requiring that it intervene in *no* scenario in which its intervention is unnecessary is certainly overkill when all that Frankfurt's argument calls for is its nonintervention in the actual world. Perhaps then the reason Frankfurt built such overkill into his counterexample is that he thought Jones's moral responsibility for murdering Smith would be clearer to the reader if the mechanism were set up so as to intervene in the bare minimum of cases required by the murder's inevitability. But whether or not BSJ, so qualified, would enjoy this rhetorical advantage, it certainly enjoys no logical advantage. Since the requirement that Black's mechanism intervene only when necessary turns out to raise unanticipated difficulties for Frankfurt's argument (of the kind set forth at the beginning of this section), the cleanest response is simply to jettison the requirement.

Suppose then that Black doesn't care in the least whether he gives himself away. His sole objective is to force Jones, *via* the "thought control" exercised by the mechanism, to murder Smith. Intervention is not contingent on mental "triggers" detected by the monitoring device; perhaps the mechanism doesn't even have a monitoring component. It is simply set to override Jones's mental processes and bring about his murder of Smith. But suppose also that the mechanism is susceptible to interference from the controlee's brain waves, and that Jones's actual thought processes set up the precise pattern of waves necessary to thwart the mechanism and keep it thwarted during the entire time it takes him to entertain, deliberate, decide, and commit the murder of Smith. (Say it's just an astounding coincidence.) If Jones had deviated even the slightest from his actual course of (mental and physical) action, the mechanism would have re-established control and achieved its normal coercive effect. But as matters actually stand the mechanism contributes nothing to the outcome.

With this variation of BSJ on hand, let us call the original counterexample 'BSJ_1' and the new one, which drops Black's desire not to show his hand, 'BSJ_2.' If Jones had no alternative to murdering Smith in BSJ_1, he had none in BSJ_2. (BSJ_2 contains all the blocked alternatives of BSJ_1 plus a bunch more.) Also, whatever

it was about BSJ_1 that made Jones seem prima facie morally responsible for Smith's murder is also present in BSJ_2: the mechanism was inoperative; it contributed nothing to Jones's action; Jones would have acted as he did if the mechanism had not even existed. The difference between the two cases is that in BSJ_1 Jones's prima facie moral responsibility is threatened by the nomological assumptions that have to be made if the mechanism is to intervene in the bare minimum of cases, whereas in BSJ_2 there is no such requirement and no such threat. If we make Frankfurt's case against PAP rest on BSJ_2, we can avoid the objection canvassed in this section, which only diverts attention away from the real issues.

A SECOND SET OF OBJECTIONS

The new counterexample makes one kind of objection irrelevant, but leaves in place all those objections which argue that Jones has sufficient alternate possibilities to subsidize his moral responsibility for Smith's murder despite the presence of the mechanism. The posited alternatives are not alternatives to Jones's killing Smith, to be sure – with few exceptions, critics appear to grant the inevitability of Smith's death at Jones's hands.[18] The claim, rather, is that there are alternatives to other things that Jones does, and it is in virtue of these other alternatives that Jones is morally responsible for killing Smith. Though BSJ_2 involves tighter control over Jones's alternatives than does BSJ_1, it is no less vulnerable to this next set of objections, since the kind of alternative that figures in the objections is endemic to both versions of Frankfurt's counterexample.

There are two directions one might go in prosecuting this strategy. The first is motivated by the fact that the mechanism leaves open at least one alternative to Jones's actual uncompelled murder of Smith, namely, Jones's *compelled* murder of Smith. There are a number of ways in which this alternative has been translated into a defense of PAP against Frankfurt's attack. One way, taken by Peter Van Inwagen, is to argue that Jones's murder of Smith comprises a different event-*particular* depending on whether it comes about through Jones acting on his own or through the intervention of the mechanism. Van Inwagen maintains that the origin of an event-particular is essential to it, so that different origins entail different

events. Let 'α' stand for the actual series of Jones's mental states leading up to his murder of Smith, 'α_j' for one of the mental states constituting α, and 'e_α' for the actual murder-event in which α terminates; and let 'β_j' stand for a possible deviation from α_j, 'β' for the series of Jones's mental states which would lead up to his murder of Smith were the mechanism's intervention to be triggered by β_j, and 'e_β' for the particular murder-event in which β would terminate. Then Van Inwagen's position is that $e_\alpha \neq e_\beta$, though both event-particulars instantiate the event-universal *Jones's murdering Smith at* t_1. Van Inwagen's defense of PAP concludes with the proposal that we are morally responsible for event-particulars, not for event-universals, and that Jones is responsible for bringing about the event-particular e_α in virtue of the availability of the alternative event-particular e_β (not to mention alternatives e_γ, e_δ, e_ε, etc.).[19]

William Rowe suggests another way the compelled alternative might be employed in defending PAP. Drawing on the libertarian notion of "agent causation," Rowe argues that if Jones is morally responsible for murdering Smith, it is because he *agent-causes* his volition to murder Smith. In the alternative scenario, on the other hand, Jones does not agent-cause his volition to murder Smith; it is the mechanism that causes his volition. There is therefore an alternative to Jones's agent-causing the murder, namely, Jones's *not* agent-causing the murder. And this is precisely the kind of alternate possibility that is relevant to Jones's moral responsibility. People are morally responsible, in the first instance, for what they directly bring about – i.e., for those mental events (including volitions) which are the first effects of their agency. It is in virtue of Jones's culpability for agent-causing a volition to murder Smith that he is culpable for murdering Smith (when this follows in the right way from his volition). Jones's primary obligation is to refrain from agent-causing a volition to murder Smith; he does not have a moral obligation to agent-cause some other volition instead of this one. His obligation not to agent-cause a volition to murder Smith is discharged when the mechanism pre-empts his powers of agent-causation. Since this alternative is available to him (in BSJ_2 as well as BSJ_1), he is morally responsible for murdering Smith under the terms of PAP.[20]

So much for the first direction in which one might look for alternatives supporting PAP. In addition to the Rowe-Van Inwagen

approach, which finds the relevant alternative in the possibility of Jones murdering Smith under coercion, there is also the approach which looks to the possibility of Jones's *un*coerced deviation from the actual sequence. BSJ_2, no less than BSJ_1, requires a trigger – its operation is contingent upon Jones deviating in some way from the actual sequence. Initial deviations are no more problematic than they would be if the mechanism didn't even exist; it is only subsequent to the deviation that the mechanism comes into play to take control of Jones's mental processes. An efficient mechanism might squelch Jones's liberty almost instantaneously; but no matter how efficiently it may operate, intervention that is contingent upon a trigger cannot bring it about that the trigger never occurred. Since the mechanism cannot affect the original deviation, there is no less reason in this case than there would be in a normal mechanism-less case to suppose the deviation itself to be uncoerced. But then this is the alternate possibility required by PAP. Jones may not be able to deviate from killing Smith, but at any point along his path toward murdering Smith Jones can deviate from what he actually does (if only for a moment, before the mechanism takes over), and these possible deviations are sufficient to ground his moral responsibility.

How successful are these two defenses of PAP – the one that looks to the coerced alternative, and the one that looks to the initial deviation, for the alternate possibility required by PAP? The former is saddled with serious, perhaps even fatal, problems. Van Inwagen's version, in requiring only that a different event-particular be available to ground ascriptions of moral responsibility, appears to trivialize PAP. It is hard to see what work this condition is doing, or under what circumstances it would fail to be satisfied, if its satisfaction requires nothing more than the possibility of a different event-particular. And there is a peculiarity that Van Inwagen's version shares with Rowe's version, namely, that the alternate possibility it offers to ground Jones's actual freedom and moral responsibility is one in which Jones is neither free nor responsible because he is coerced by the mechanism. As John Martin Fischer notes, "it would be very puzzling and unnatural to suppose that it is the existence of various alternative pathways along which one does *not* act freely that shows that one has control of the kind in question."[21] Insofar as the coerced alternative approach is at

all plausible as a defense of PAP, it is surely because the coerced alternative originates in an *un*coerced deviation from the actual sequence. This makes it parasitic on the initial deviation approach.

Is the latter approach any more successful? This depends on whether the deviations available to Jones, in the split second before an efficient mechanism could assert control over the situation, are of the right sort to satisfy PAP. John Fischer maintains that they are not. Fischer assimilates both the coerced alternative and the initial deviation approaches to what he calls the "flicker of freedom strategy," and argues that both fail.[22] The basic problem with "flicker" strategies, as Fischer characterizes them, is that the alternate possibilities they offer are insufficiently robust to ground moral responsibility.

It is easy to see why the "flicker" epithet might be justified in the case of the coerced alternative approach. The mere ability to bring about a different event-particular is pretty thin gruel for moral responsibility to subsist on. Likewise the availability of volitions that are not agent-caused is rather disappointing as an alternate possibility (compared with the agent's ability to agent-cause alternative volitions). But why suppose that the initial deviation approach is open to the same objection? Most deviations are doubtless too trivial to constitute the alternate possibilities required by moral responsibility. (This is especially clear in BSJ_2, where any deviation at all is sufficient to trigger the mechanism.) But the triviality of most deviations is insufficient for Fischer's critique: PAP requires only that there be *at least one* deviation open to Jones which is sufficiently robust to constitute the alternate possibility entailed by moral responsibility, and it's hard to see how the existence of at least one such deviation can be ruled out in advance. Fischer's response to this concern is to reconstruct Frankfurt's example so that every deviation which would be sufficiently robust to support PAP is preceded by some involuntary sign or indication, like a blush or a twitch. In this case Black could intervene after the involuntary sign in order to preclude the voluntary deviation, leaving the sign itself as the only deviation. "Here the 'triggering event' (i.e., what would trigger the intervention of Black) is *not* any sort of initiating action, and thus cannot be said to be freely done," Fischer writes. As a result, "this sort of triggering event appears to be not sufficiently robust

to ground responsibility ascriptions."[23] As we noted in the previous section, however, the positing of such signs or indications may beg the question against libertarianism.

In looking more closely at the initial deviation approach, there are really two kinds of cases to be considered. The first is where the deviation comes only at the last moment, with Jones simply refraining from murdering Smith. And why shouldn't this be possible? After all, the mechanism is inoperative, given the actual sequence; it would intervene only if the sequence deviated in some way. But the intervention can only affect what happens *after* the original deviation – the deviation itself cannot be affected. So suppose that S saves his deviation until t, when he simply refrains from A-ing. There is nothing that the mechanism (even the one in BSJ_2) can do to prevent this. In sum, Frankfurt pretends to describe a mechanism which, assuming an actual sequence terminating in an action A at time t, can *guarantee* that the sequence will so terminate, despite the ever-present possibility of an initial deviation from the sequence. But this description is incoherent, given the possibility that the deviation might come *at t*.

There are two responses available to the critic which make it very doubtful that this defense of PAP can succeed. The first begins by asking what would happen if Jones were to avoid murdering Smith at t by reserving his initial deviation until then. The answer, of course, is that the deviation would trigger the mechanism, which would then compel Jones to murder Smith at t + i (a split second after t). The alternative to Jones's uncoerced murder of Smith at t is apparently his coerced murder of Smith at t + i. As a candidate for the alternate possibility required by PAP, this alternative is open to Fischer's "flicker of freedom" critique. The situation can certainly be elaborated in ways that strengthen it against Fischer's critique. We might imagine, for example, that Jones's window of opportunity for killing Smith closes at t (when he is himself killed by police sharpshooters deployed to defend Smith against assassination), so that it is no longer possible at t + i for the mechanism to bring about Smith's death through control over Jones's mental processes. Here the alternative is not a minimally different killing of Smith, but a deliberate, uncoerced, and final avoidance of killing Smith, which is arguably more than a mere "flicker of freedom." It is important

to remember, however, that the critic of PAP is not obliged to show that in *every* case of moral responsibility there are no morally relevant alternatives; he is only required to show that in *some* case(s) there are none. The critic should therefore reply that in the specified counterexample the window of opportunity remains open; Jones's opportunities in other scenarios are beside the point.

The second problem with this defense of PAP is that it depends crucially on the nature of A-ing: what it is to A and what it would be to refrain from A-ing. Not every action is such that abstaining from it could even *be* an initial deviation, since an initial deviation lasts only long enough for the mechanism to squelch it, which might be a vanishingly short time. So long as A-ing/not-A-ing has any temporal complexity, requiring more than a split second to bring to completion, an efficient mechanism can nip any deviation in the bud. What this defense presupposes, then, is an action that is temporally simple. But then it is insufficient to rehabilitate PAP as a general stricture on any case of moral responsibility.

So it seems that Frankfurt and Fischer can construct a counterexample involving a complex action such that satisfaction of PAP could not possibly rest on delaying the initial deviation until t. (BSJ is itself such a counterexample, since actions which include bodily movements – e.g., firing a gun – are always complex.)[24] There remains to consider, then, the second type of case, where PAP is satisfied in virtue of initial deviations available *earlier* in the sequence. This case is more favorable to the defenders of PAP: while critics can certainly specify a counterexample in which not-A-ing is complex and therefore preventable by the mechanism, the sequence of mental and bodily events (and possible deviations therefrom) leading up to S's A-ing at t will surely include some simple actions as well. Since these simple actions cannot be controlled by the mechanism in BSJ_2 (not to mention BSJ_1), the only remaining question is the relevance to S's moral responsibility of the fact that these (simple) alternatives are available.

Is there reason to think that every complex action for which one is morally responsible is preceded by some simple action from which it is possible to deviate instantaneously and in virtue of which possible deviation one is morally responsible for the later action? David Widerker has argued recently that there is a simple

action which satisfies all these requirements, namely, the *decision* to A (not-A) at t. In the first place, Widerker maintains, deciding is simple:

> it would be conceptually wrong for one to describe what Jones is doing at a given moment by saying that he is in the *process* of deciding to kill Smith, or that he has not yet finished deciding to kill Smith. Jones, to be sure, can be said to be in the process of *trying to reach a decision* whether to kill Smith or not. But that process and the event of deciding to kill Smith are two different things.[25]

Because deciding is a simple mental action, the mechanisms in BSJ_1 and BSJ_2 cannot control it. In the second place, deciding (or something like it) always accompanies morally responsible actions; so it is impossible for Frankfurt *et al.* to come up with a counterexample to PAP which does not involve decision. And finally, it is S's control over deciding to A, including his power to decide not to A, which makes him morally responsible for A-ing. So PAP, at least in the broader reading defined by PAP+, survives any possible counterexample employing a BSJ-type mechanism.

This is the most plausible ground on which to attack BSJ, but it is still far from decisive. For one thing, it's not clear that every action for which a person is morally responsible is preceded by a simple decision of the type Widerker describes. We hold people responsible for what they do habitually, when no identifiable decision is present; and some nonhabitual actions which are candidates for moral praise or blame arise from intentions which are formed only gradually, without a single shining moment of decision. What's more, even cases fitting Widerker's decision-model remain vulnerable to counterfactual intervention under the right circumstances. Suppose that (i) there are various necessary conditions that must be satisfied if Jones is to alter his intention to murder Smith; (ii) at least one of these necessary conditions is absent from the actual sequence; and (iii) this same necessary condition is one of the triggers for Black's mechanism. Then the situation is one in which the alternative decision required by PAP cannot be reached *via* a single (original) deviation from the actual sequence, and the decision is therefore preventable by the mechanism in BSJ_2.

A likely reply to this objection is that a case meeting the foregoing specifications, if offered as a counterexample to PAP, begs the question against a libertarian like Widerker. If any condition

necessary for Jones's deciding otherwise is missing from the actual sequence, then actual conditions are *sufficient* for Jones's deciding as he does; and if this is so, the libertarian won't accept the case as one in which Jones is morally responsible despite the unavoidability of his decision.[26]

To see that cases of this type needn't beg the question, consider the following way that (i)–(iii) might be satisfied. Jones is a hardened criminal, but not so hardened that he isn't morally responsible for his criminal deeds. Absent the counterfactual intervener, Jones *could* decide against murdering Smith; but murder is the path of least resistance for him, and deviating from it would require a certain mental effort. The requisite effort might involve a moral struggle, a prudential calculation, a full-scale review of his life and priorities, the rereading of religious texts that had been important to him as a child, and so on. Jones *can* do these tings, and through them (possibly) arrive at a decision not to murder Smith; but he can't reach this decision from his present course in one simple mental act. And this is all the more true the farther along he is on that course: the suggestion that Jones (given who he is) might be well advanced in his deliberations, with everything pointing toward a decision to commit murder, and then suddenly perform a simple about-face, without first entertaining doubts about his former course or passing through any intermediate step (however brief), looks like an appeal to *magic*. Nevertheless, Jones is surely morally responsible for his decision and for the action to which it leads, and the libertarian should not be constrained to say otherwise; after all, at any point prior to the actual decision the opposite decision was accessible to Jones, even though it was not *immediately* accessible. Given this appraisal of the situation in the absence of outside interference, suppose now that Black's mechanism is introduced into the picture, and that it is programmed so that any of the intermediate steps by which Jones might reach a decision not to murder Smith would trigger its intervention. Then Jones can't refrain from deciding to murder Smith; but intuitively, he is still responsible for that decision, since he made it on his own and the mechanism had no effect on the actual sequence leading up to the decision.

While the foregoing may provide an adequate counter to Widerker's argument, it is hardly the last word on the subject.

For one thing, though it avoids begging the question against libertarianism *per se*, it still begs the question against that species of libertarianism – "*Sartrean* libertarianism," as we might call it – which insists on the perpetual possibility of 180-degree turns in our life as a requirement of robust free will. While this may not be an otherwise attractive position, it does represent an undefeated possibility for defending PAP. For another thing, Jones still has available to him plenty of alternatives, other than the possibility of deciding otherwise, and there is no guarantee that the pro-PAP forces won't regroup around one of these – e.g., Jones's power, in the split-second before the mechanism intervenes, to entertain the shadow of a doubt about his plans for murder. This is still a "flicker of freedom." Extinguishing it altogether, by selecting a purely physiological state (or other morally irrelevant factor) as the intervention-triggering necessary condition for Jones's deciding otherwise, would reintroduce worries about causal determinism, leaving the libertarian (and not just the Sartrean one this time) unimpressed. The main advantage of the anti-PAP position – that it need produce only a single counterexample – is matched by the pro-PAP advantage, which is that it need identify only a single (potentially relevant) alternative. So it's hard to see how a definitive resolution of the issue can be found in this direction.[27]

COUNTEREXAMPLES WITHOUT COUNTERFACTUAL INTERVENTION

While it may be premature to declare a stalemate in the current discussion, it is nevertheless worth considering whether some alternative approach might yield greater progress. From the anti-PAPist perspective, I should think, the best response to problems with Frankfurt's counterexample would be to dispense with the feature that makes it repeatedly vulnerable to attack. The fundamental problem with BSJ is that the unavoidability of the murder rests on a mechanism employing *counterfactual* intervention, which by its very nature permits initial deviations. This feature of Frankfurt's original example is not essential to a Frankfurt scenario *per se*. Frankfurt's stipulation that the mechanism would have exercised its coercive function only if things had gone differently than they

actually did was simply the means by which he ensured that the conditions making for unavoidability would play no actual role in bringing about Jones's murder of Smith. But why suppose that the only way to ensure that $C_N \neq C_\alpha$ is *via* a coercive mechanism that operates counterfactually?

The original "Frankfurt scenario" is of course John Locke's famous example in which "a man be carried whilst fast asleep into a room where is a person he longs to see and speak with, and be there locked fast in, beyond his power to get out; he awakes and is glad to find himself in so desirable company, which he stays willingly in, i.e. prefers his stay to going away."[28] In this case the man is morally responsible for what he does even though he cannot do otherwise. (Say he has a philosophy class to teach, and he knows that remaining in the room will make it impossible for him to get to class.)[29] But the condition that makes for unavoidability in this example is not counterfactual in nature: the door is actually locked; it doesn't lock only when someone approaches the door and tries to leave. What makes the locked door compatible with the man's moral responsibility is simply that it is not among the conditions actually leading the man to stay in the room. It disengages itself from those conditions, not by retreating to a set of nearby possible worlds (access to which requires a counterfactual trigger), but by waiting on the sidelines in the actual world. Of course a locked door still provides an agent with plenty of alternatives, so it might be doubted how far Locke's example goes toward refuting PAP, particularly when understood along the lines of PAP+. (The man could still *try* to leave, and the availability of this alternative is surely enough to undergird his moral responsibility.) But Locke's example does provide some initial encouragement that the unavoidability essential to a Frankfurt scenario does not have to rest on a counterfactual device.

Thus encouraged, let me suggest three different Frankfurt scenarios in which unavoidability does not wait upon a counterfactual trigger and so can extend to all the agent's actions, leaving no alternate possibilities to ground moral responsibility. The first is a development of Locke's example. The basic principle that immunizes moral responsibility against Locke's locked door does not appear to depend on the amount of "elbow room" between the agent's will and external obstacles to alternative courses of action;

it has to do solely with the fact that these obstacles are not part of the sequence of states actually productive of the agent's behavior. (It wouldn't alter our ascription of moral responsibility one iota if we learned that the locked door was right at the man's elbow rather than on the far side of the room, or that the room itself was 50 sq. ft. rather than 500 sq. ft.) What is needed is a case in which unavoidability arises solely from blocked alternatives (rather than compulsion in the actual sequence), and the amount of elbow room remaining to the agent approaches zero. Imagine then a mechanism that blocks neural pathways rather than doorways. Suppose that the actual series of Jones's mental states leading up to the murder of Smith is compatible with $\overline{\text{PAP}}$, except that the mechanism is in operation. The mechanism is not intervening directly in the series itself; it is allowing the series to unfold on its own, but simply blocking all alternatives to the series. Of course it can't block alternatives *in response* to the way the series is unfolding, because then the blockage would be coming too late to have any effect on the avoidability or unavoidability of Jones's actions. Instead, the mechanism blocks alternatives in advance, but owing to a fantastic coincidence the pathways it blocks just happen to be all the ones that will be unactualized in any case, while the single pathway that remains unblocked is precisely the route the man's thoughts would be following anyway (if all neutral pathways were unblocked). Under these conditions, the man appears to remain responsible for his thoughts and actions, given the same intuitions at work in (MI).[30]

The second example of a Frankfurt scenario which eschews counterfactual compulsion is a variation on the suggestion, made by David Blumenfeld and John Fischer, that actions with significant implications for Jones's killing or not killing Smith might be preceded by a sign or indicator. The latter were understood to be such that, given a particular indicator, the action of which it is the indicator would certainly follow unless prevented by Black's mechanism. Here the unavoidability of Jones's murder of Smith rested on the fact that indicators for actions at variance with the murder would trigger the mechanism to intervene. This strategy is open to potential objections, of the sort canvassed earlier. Suppose, however, that the counterfactual mechanism is omitted from the picture, leaving only the indicators for Jones's actual action. Since

these indicators guarantee that the action will take place, their occurrence is itself sufficient to render the action unavoidable. Of course, if Jones's action is unavoidable because it is *causally determined* by its indicator(s), this will beg the question against libertarianism. So posit an actual sequence satisfying $\overline{\text{PAP}}$ (including causal indeteminism and/or agent-causation at all those junctures where libertarians would require it for Jones's moral responsibility), with the further stipulation that the sequence is continuously emitting tachyonic signals (consisting of particles which travel backward in time) encoding complete and unambiguous information about Jones's current states. Then Jones's killing of Smith, along with all the preliminaries to it (such as deciding to kill Smith) that are relevant to his moral responsibility, will be preceded by tachyonic indicators given which Jones's actions are unavoidable. But these indicators do not causally determine Jones's future action(s); they are not even part of the actual sequence leading up to the murder (they come "later" in the sequence – not temporally, but explanatorily). When Jones's murder of Smith is placed in this context, his responsibility for the murder is sanctioned by the same moral intuitions captured in (MI).

For a final example of moral responsibility combined with non-counterfactual inevitability, imagine that Black is the predictor from Newcomb's puzzle, and that his perfect track record of predictions reflects the fact that he is not just inerrant (so far) but *essentially* inerrant. As in BSJ, Jones is going to murder Smith at t, and he is going to do so on his own. But unlike the set-up in BSJ, Black secures the unavoidability of the murder, not by wielding an intrusive device, but simply by predicting that Jones will murder Smith at t. Given Black's infallibility, once the prediction is made Jones cannot do otherwise than murder Smith at t; if we add that the prediction was made before Jones was even born, we can further conclude that there is *no* time at which Jones has it within his power to refrain from murdering Smith. (For Jones's action to be unavoidable in this sense, it is enough that Black's infallibility rest on the physical impossibility of error; it is not necessary that he should also be logically or metaphysically inerrant.) Jones's moral responsibility in the circumstances might still rest on some deviation that he has the power to make prior to murdering Smith; but this possibility

can be eliminated simply by adding that the predictor is essentially *omniscient* as well as infallible, rendering the entire sequence of Jones's actions (mental and physical) unavoidable. Once again (MI) supports the judgment that Jones is morally responsible for murdering Smith.

In each of these three cases we have the same reason for thinking that Jones is nevertheless morally responsible as we had in BSJ. Frankfurt himself characterized the exculpatory features of BSJ this way: the mechanism "played no role at all in leading [Jones] to act as he did"; indeed, "everything happened just as it would have happened without Black's presence in the situation and without his readiness to intrude into it"; for this reason the counterfactual intervener is "irrelevant to the problem of accounting for a person's action" and "does not help in any way to understand either what made [Jones] act as he did or what, in other circumstances, he might have done."[31] But if this is true in BSJ, it is all the more true in the new cases, which dispense with counterfactual mechanisms. Infallible beliefs about the future, for example, rule out alternate possibilities without their making any difference to, or playing any role in, or helping in any way to explain the future. They therefore constitute a violation of PAP under the "master intuition" formulated in (MI). And the same appears to be true in the other two cases as well.[32]

CONCLUSION

The three counterexamples in the preceding section satisfy all the conditions for a Frankfurt scenario. Each provides a case in which, by the lights of (MI) – a principle that Frankfurt's critics appear to accept, or at least prefer to leave unchallenged – Jones is morally responsible for his actions despite their unavoidability. But none of these cases is open to the objections levelled against Frankfurt's original counterexample, all of which turn on the counterfactual intervention by which the unavoidability of Jones's actions is secured.

There is one respect, however, in which the new counterexamples are more vulnerable than the original. We are pretty confident that BSJ is possible (logically, metaphysically, even physically). We are

probably less confident that the new counterexamples are possible (we may even be pretty confident that some or all of them are impossible). The first counterexample is supposed to differ only in degree, not in kind, from Locke's locked room example; in particular, it presupposes that alternative pathways can be blocked without this affecting what's going on in the actual pathway, even when the actual pathway is the *only* one that remains unblocked. But this presupposition might not survive examination. (Maybe it would; but maybe it wouldn't.) Likewise, the second counterexample presupposes the intelligibility of reverse causation, while the third presupposes the possibility of infallible foreknowledge; but these are hardly uncontroversial presuppositions.

It is doubtful, however, that this liability is enough to nullify the new counterexamples' effectiveness. In the first place, the claim that all three of the counterexamples are incoherent is at least as controversial as the claim that at least one of them is conceptually sound. So if PAP is true only if all the counterexamples fail, its truth is infected with the same degree of controversy. At the very least, then, these counterexamples effectively undermine the idea, ascribed by Frankfurt to "some philosophers," that PAP is so firm, unquestionable, and overwhelmingly plausible that it might even rise to the level of an a priori truth.[33] Secondly, even in the worst case where all three of the counterexamples turn out to rest on some conceptual confusion, the confusion is not right on the surface. This makes it doubtful that the confusion plays any role in our reflections on the counterexamples or in the conclusions they help us reach regarding PAP. The counterexamples can continue to give useful and illuminating form to intuitions that run counter to PAP, even if there is some hidden incoherence in the notion of retrocausation and the other unavoidability devices employed in the counterexamples. If the minimal changes needed to purge the counterexamples of logical taint were to reintroduce a measure of avoidability, there would still be little reason to think that Jones's moral responsibility depends on the reinstatement of these alternatives. Of course the assumption that all three counterexamples might *be* logically flawed is one that I am making here only for the sake of argument.

Supposing that this defense of the counterexamples is acceptable, it is nevertheless noteworthy that none looks much like the

ordinary conditions under which people act. Even if the counterexamples succeed in breaking the conceptual connection between moral responsibility and alternate possibilities, it is not immediately clear what actual difference this should make to classic debates over moral responsibility and freedom of the will, which aim at clarifying the human situation. I conclude with brief comments on two of these debates.

One of the oldest threats to moral responsibility on grounds of inevitability is *fatalism*. Unadorned "logical" or "prior-truth" fatalism, from the variety formulated in chapter 9 of Aristotle's *De Interpretatione* to the one defended in Richard Taylor's *Metaphysics*,[34] appears to rest on a modal fallacy, and consequently does not pose a live problem for PAP. But "theological" fatalism, in which inevitability stems from the foreknowledge of an infallibly omniscient being, is not so easily dismissed.[35] The problem of theological fatalism does appear to dissolve, however, given the case against PAP based on our counterexamples. Divine foreknowledge may eliminate alternatives to the actual sequence of events, but inasmuch as it does so without causing, explaining, or making any difference to the actual sequence, the unavoidability with which it endows future actions cannot detract from moral responsibility for those conditions.[36]

An equally ancient concern, which has come to eclipse the problem of theological fatalism in modern discussions of moral responsibility, is the threat posed by *causal determinism*. This threat seems no less real after consideration of the counterexamples has led us to reject PAP. (MI) permits moral responsibility when the conditions making for an action's unavoidability are distinct from the conditions actually giving rise to that action. But the prior events that causally determine the action, if determinism is true, belong to both sets of conditions. So there is nothing in (MI) itself that should lead us to regard causal determinism as compatible with moral responsibility. Though Frankfurt and Fischer are both inclined to think that the defeat of PAP paves the way for a morally respectable compatibilism, this conclusion goes beyond what is actually sanctioned by the argument. Fischer in particular argues that the only reason causal determinism seems incompatible with moral responsibility is its exclusion of alternate possibilities, and that once PAP is rejected

and unavoidability is no longer regarded as a threat to responsibility there is no further reason to regard determinism as a threat either.[37] This implies that it is irrelevant to moral responsibility whether $C_N = C_\alpha$, and it's not clear that this is so. Certainly it makes a difference to (MI), and through it to the judgments that we make regarding counterexamples to PAP. Compatibilism would require considerable further argument – perhaps no less argument after the defeat of PAP than before.[38] The threat posed by causal determinism, unlike the one posed by divine foreknowledge, does not appear to vanish in the face of Frankfurt's argument against PAP.

But this controversy lies beyond the bounds of the present paper, whose main conclusion can now be summarized. The fact that critics share Frankfurt's fundamental intuition in favor of (MI) means that problems with BSJ are not necessarily fatal to Frankfurt's case against PAP. These problems ultimately stem from the counterfactual character of the mechanism by which BSJ enforces unavoidability while preserving moral responsibility. But counterfactual mechanisms are not the only way to generate the scenario which is the focus of (MI) – I have suggested three others. This is enough, in my view, to break the conceptual connection between moral responsibility and alternate possibilities. At the very least, the ball is back in the critics' court to show how one can accept (MI) while rejecting Frankfurt's case against PAP.[39]

NOTES

[1] Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66 (December 4, 1969), pp. 829–839.

[2] *Ibid.*, p. 829.

[3] Frankfurt's challenge is contained in "Alternate Possibilities and Moral Responsibility," where he notes (p. 835, n. 2) that a similar argument had been formulated in lectures by Nozick. For Fischer's contributions, see especially his "Responsibility and Control," *Journal of Philosophy* 79 (January 1982), pp. 24–40, and recent book, *The Metaphysics of Free Will: An Essay on Control*, Aristotelian Society Series, v. 14 (Oxford, UK & Cambridge, Mass.: Blackwell, 1994).

[4] David Blumenfeld, "The Principle of Alternate Possibilities," *Journal of Philosophy* 68 (June 3, 1971), p. 339.

[5] "Alternate Possibilities and Moral Responsibility," p. 830.

[6] *Ibid.*

[7] *Ibid.*, p. 835.

[8] This is why James W. Lamb appears to me to achieve at most a hollow victory against Frankfurt *et al.* in his "Evaluative Compatibilism and the Principle of Alternate Possibilities," *Journal of Philosophy* 90 (October 1993), pp. 517–527. His "weak principle of alternate possibilities" is essentially PAP+ minus its parenthetical clause: "the thesis that a person is morally responsible for doing something only if at some time there is *something* he could have avoided doing" (p. 527). So far as this principle goes, Jones might be morally responsible for murdering Smith even though he faced only one alternate possibility in his entire life – say, at his first birthday party when he sucked on the black jelly bean but could have sucked on the pink one instead. But if Jones is morally responsible for murdering Smith given this single alternate possibility, he is surely morally responsible for murdering Smith when this possibility is subtracted from the situation – because it can't possibly be relevant to his responsibility for murdering Smith that he could have sucked a differently colored jelly bean on his first birthday. Though Lamb's weak principle of alternate possibilities seems so weak that it can't possibly sustain the connection between alternate possibilities and moral responsibility, it is worth noting that the counterexamples I ultimately develop against PAP are equally damaging to his weak principle as well.

[9] This amendment is suggested in Blumenfeld, *op. cit.*, and in Robert Cummins, "Could Have Done Otherwise," *Personalist* 60 (October 1979), pp. 411–414. Here, and in the principle referenced in footnote 10, I have supplied the contrapositive of the formula actually employed by the authors, in order to make the wording more comparable with that of PAP.

[10] This is Frankfurt's own suggestion, in "Alternate Possibilities and Moral Responsibility," pp. 838–839, for what remains true once PAP is abandoned. See footnote 9.

[11] Since the content of $\overline{\text{PAP}}$ gets filled in differently by different theories of moral responsibility, this part of the Frankfurt scenario must be left suitably vague, lest Frankfurt's argument defeat PAP for one theory but not for others. Except for its tolerance of unavoidability, a good Frankfurt scenario should be neutral between competing theories of moral responsibility.

[12] Frankfurt himself, in a more recent essay, characterizes what I am calling a 'Frankfurt scenario' in the following terms: "The distinctively potent element in this sort of counterexample to PAP is a certain kind of overdetermination, which involves a sequential fail-safe arrangement such that one causally sufficient factor functions exclusively as backup for another. The arrangement ensures that a certain effect will be brought about by one or the other of the two casual factors, but not by both together. Thus the backup factor may contribute nothing whatever to bringing about the effect whose occurrence it guarantees." See "What We Are Morally Responsible For," in *How Many Question? Essays in Honor of Sidney Morgenbesser*, ed. Leigh S. Cauman, Isaac Levi, Charles Parsons, and Robert Schwartz (Indianapolis: Hackett, 1982).

[13] "Alternate Possibilities and Moral Responsibility," p. 837.

[14] David Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities," *Philosophical Review* 104 (April 1995), pp. 248–249.

[15] *Ibid.*, p. 251.

[16] David Widerker offers a similar argument in "Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities."

[17] Widerker, "Libertarian Freedom and the Avoidability of Decisions," and Fischer, "Libertarianism and Avoidability: A Reply to Widerker," both in *Faith and Philosophy* 12 (January 1995), pp. 113–125. I discuss this exchange in "Frankfurt Counterexamples: Some Comments on the Widerker-Fischer Debate," *Faith and Philosophy* 13 (July 1996), pp. 395–401.

[18] Robert Heinamann, in "Incompatibilism without the Principle of Alternative Possibilities," *Australasian Journal of Philosophy* 64 (September 1986), pp. 266–276, may be an exception – but it's not clear that the sense of evitability and inevitability at work in his argument is the one that is relevant to moral responsibility.

[19] Peter van Inwagen, "Ability and Responsibility," *Philosophical Review* 87 (April 1978), pp. 201–224.

[20] William L. Rowe, "Two Concepts of Freedom," *Proceedings and Addresses of the American Philosophical Association* 61 (1987), pp. 43–64. A similar approach, which does not appeal to agent-causation but still fits the general contours of Rowe's line, is taken by Margery Bedford Naylor in "Frankfurt on the Principle of Alternate Possibilities," *Philosophical Studies* 46 (September 1984), pp. 249–258. According to Naylor, "Frankfurt's case is one in which it is entirely up to Jones whether or not to do A on his own but not entirely up to him whether or not to do A" (p. 252). Naylor argues that this is compatible with Jones's moral responsibility under the terms of PAP. The reason is that Jones is responsible, not for murdering Smith, but for *murdering (choosing to murder) Smith on his own*. The latter is not unavoidable: were Jones to trigger the mechanism he would thereby gain access to an alternative possibility, namely, *murdering (choosing to murder) Smith under coercion from the mechanism*. This is the morally relevant alternative, and its availability shows PAP to be undefeated by Frankfurt's counterargument.

[21] *The Metaphysics of Free Will*, p. 141.

[22] *Ibid.*, pp. 134–147.

[23] *Ibid.*, p. 144.

[24] Or change the example to make the complexity more evident. Suppose Jones has an obligation to make his child-support payments by t, but he deliberately refrains from doing so; nor does he at any point so much as take any steps in the direction of making the payments. Because of its complexity, any deviation toward making the child-support payments could be nipped in the bud by the mechanism, even if the deviation were reserved until the last moment. Yet Jones is morally responsible for his failure to pay child-support by t.

[25] "Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities," p. 253.

[26] I want to thank Bill Rowe for bringing this objection to my attention.

[27] Fischer himself, in *The Metaphysics of Free Will*, acknowledges the difficulty of defeating a determined PAPist: "a flicker theorist can point out that even the

fanciest, most sophisticated Frankfurt-type example contains *some* alternative possibility, no matter how exiguous. And, indeed, it is hard to imagine how to construct any kind of non-question-begging example in which it is clear both that there are absolutely no such possibilities and the agent is morally responsible for his action" (p. 145).

[28] John Locke, *An Essay Concerning Human Understanding* Bk. II, ch. XXI, §10. Of course Locke has his own agenda here, which is not entirely the same as Frankfurt's.

[29] Or suppose that he finds the other person's company desirable because he knows that his presence is guaranteed to annoy that person. Michael J. Zimmerman, in "Moral Responsibility, Freedom, and Alternate Possibilities," *Pacific Philosophical Quarterly* 63 (1982), pp. 243–254, offers this embellishment on Locke's example, adding that "Harry Frankfurt ... bases his argument on cases essentially similar to the case just given" (p. 243).

[30] I develop this example at greater length in my "Freedom, Foreknowledge, and Frankfurt" (under consideration), presented in 1996 at the Central Division Meeting of the American Philosophical Association.

[31] "Alternate Possibilities and Moral Responsibility," pp. 836–837.

[32] On pp. 146–147 of *The Metaphysics of Free Will*, Fischer appears to suppose that the *only* way to rule out absolutely all alternatives to an action is by causally determining that action: "if we simply imagined that *all* the alternative possibilities disappear by positing the truth of causal determinism (together with incompatibilism about causal determinism and alternative possibilities), we appear to beg the issue to which the claim about the lack of a requirement of alternative possibilities for moral responsibility was designed to apply. Thus, the sort of metaphysical gridlock characteristic of Dialectical Stalemates again rears its ugly head." The three examples in this section, by avoiding causal determinism in the actual sequence, are supposed to break the stalemate.

[33] *Ibid.*, p. 829.

[34] See the chapter "Fate" in Taylor's *Metaphysics* (Englewood Cliffs, N.J.: Prentice-Hall, 1963).

[35] For the superiority of theological over logical fatalism, see David Widerker, "Two Forms of Fatalism," in *God, Foreknowledge and Freedom*, ed. J.M. Fischer (Stanford: Stanford U. Press, 1989).

[36] This is St. Augustine's solution to the problem of theological fatalism, as I argue in my "Augustine on Theological Fatalism: The Argument of *De Libero Arbitrio* III.1–4," *Medieval Philosophy and Theology* 5 (1996), pp. 1–30. I defend its title to be the correct solution to this problem in my "Freedom, Foreknowledge, and Frankfurt," *op. cit.*, and in my "On Augustine's Way Out," *Faith and Philosophy* (forthcoming). Finally, I challenge the relevance of purely theological solutions in my "What *Is* the Problem of Theological Fatalism?" *International Philosophical Quarterly* (March 1998), pp. 17–30.

[37] The full argument for this position in Fischer's *The Metaphysics of Free Will* is considerably richer than I can deal with adequately here; see especially pp. 147–154. Fischer's position has shifted somewhat since his earlier "Responsibility and

Control," *Journal of Philosophy* 89 (January 1982), pp. 24–40, where he could write: "the approach I am developing concedes this dissociation [of responsibility from control], but argues that the transition from this discussion to the compatibility of determinism with responsibility is a spurious transition. This is because the reason why determinism threatens responsibility is not *that* it undermines control, but because of the way in which it undermines control; determinism involves actual-sequence compulsion, and such compulsion might be incompatible with moral responsibility" (p. 34).

[38] Much of this further argument in *The Metaphysics of Free Will* involves Fischer's account of "weak reasons-responsiveness" in ch. 8. Perhaps the rejection of PAP removes one obstacle to the compatibility of moral responsibility with a weakly reasons-responsive but causally determined actual sequence; but it by no means entails it.

[39] I am especially grateful to William Rowe, Leslie Stapp, and David Widerker for their useful comments on an earlier draft.

Department of Philosophy
Whittier College
Whittier, CA 90608, USA