Australasian Journal of Philosophy
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713659165

Actions, thought-experiments and the 'Principle of alternate possibilities'

Maria Alvarez [a]
[a] University of Southampton,

First Published:March2009

To cite this Article Alvarez, Maria(2009)'Actions, thought-experiments and the 'Principle of alternate possibilities'',Australasian Journal of Philosophy,87:1,61 — 81
To link to this Article: DOI: 10.1080/00048400802215505
URL: http://dx.doi.org/10.1080/00048400802215505

Australasian Journal of Philosophy
Vol. 87, No. 1, pp. 61–81; March 2009

Routledge
Taylor & Francis Group

ACTIONS, THOUGHT-EXPERIMENTS AND THE 'PRINCIPLE OF ALTERNATE POSSIBILITIES'

Maria Alvarez

In 1969 Harry Frankfurt published his hugely influential paper 'Alternate Possibilities and Moral Responsibility' in which he claimed to present a counterexample to the so-called 'Principle of Alternate Possibilities' ('a person is morally responsible for what he has done only if he could have done otherwise'). The success of Frankfurt-style cases as counterexamples to the Principle has been much debated since. I present an objection to these cases that, in questioning their *conceptual* cogency, undercuts many of those debates. Such cases all require a counterfactual mechanism that *could* cause an agent to perform an action that he cannot avoid performing. I argue that, given our concept of what it is for someone to act, this requirement is inconsistent.

Frankfurt-style alleged counterexamples are cases where an agent is morally responsible for an action he performs even though, the claim goes, he could not have avoided performing that action. However, it has recently been argued, e.g. by John Fischer, that a counterexample to the Principle could be a 'Fischer-style case', i.e. a case where the agent can either perform the action or do nothing else. I argue that, although Fischer-style cases do not share the conceptual flaw common to all Frankfurt-style cases, they also fail as counterexamples to the Principle.

The paper finishes with a brief discussion of the significance of the Principle of Alternate Possibilities.

I. Introduction

In his article 'Alternate Possibilities and Moral Responsibility' [Frankfurt 1969], Harry Frankfurt argued that a man may be morally responsible for an action, even though he 'could not have done otherwise'. What matters for moral responsibility, Frankfurt went on to argue in this and subsequent papers, is *why* one does what one does. In short, in that paper Frankfurt attempted to refute what he called the 'Principle of Alternate Possibilities' ('the Principle'), namely, that 'a person is morally responsible for what he has done only if he could have done otherwise' [ibid. 829]. Frankfurt argued that, far from being an '*a priori* truth' (loc. cit.), as many had previously thought, this principle is in fact false.

Frankfurt's attempted refutation of the Principle consisted in a thought-experiment which, he claimed, provides a counterexample to the Principle.

Australasian Journal of Philosophy
ISSN 0004-8402 print/ISSN 1471-6828 online
http://www.tandf.co.uk/journals DOI: 10.1080/00048400802215505

His and subsequent versions of this thought-experiment have become known as 'Frankfurt-style cases'. I shall argue that Frankfurt-style cases do not constitute counterexamples to the Principle.

The debate about the success or otherwise of such cases has gone on for over thirty years,[1] and many before me have argued that Frankfurt's thought-experiment, and generally Frankfurt-style cases, fail to falsify the Principle. Most of the critics have argued against these cases motivated by commitment to the doctrine of the incompatibility of free will/moral responsibility and determinism, and some have provided seemingly devastating objections to Frankfurt-style cases.[2] My own arguments are not motivated by any position in the debate about the compatibility of free will/moral responsibility and determinism.[3] Rather, I wish to highlight and challenge an assumption that underlies all Frankfurt-style cases and show that, given the concept of what it is for someone to *perform an action*, that assumption is untenable.

My discussion will mainly focus on Frankfurt's thought-experiment as described in his 1969 paper, rather than on the latest refined and improved versions of the example, although I shall also comment on those when appropriate. But, as I hope will become clear, my strategy is justified because my objection concerns a feature that is not only common but *crucial* to all Frankfurt-style cases.

Frankfurt presented his thought-experiment as a counterexample to the Principle because, he claimed, it is a case where an agent is morally responsible for an action he performs even though he could not have avoided performing that action. However, it has recently been argued, e.g. by John Fischer,[4] that a counterexample to the Principle need not be a Frankfurt-style case. According to Fischer, there are cases that differ significantly from Frankfurt-style cases ('Fischer-style cases') that also falsify the Principle. My main aim in this paper is to show that there is something conceptually awry with Frankfurt-style cases. Nonetheless, I believe and will try to show (Section IV.A) that Fischer-type examples do not succeed in undermining the Principle either. The paper concludes with a brief discussion of the significance of the Principle of Alternate Possibilities.

II. The Principle of Alternate Possibilities

The Principle of Alternate Possibilities is traditionally articulated in a form that Frankfurt himself uses at the beginning of his paper, namely that 'a

[1]As is well known, the volume of literature on this issue is phenomenal. For fairly exhaustive reviews see [Fischer 1999: 109 – 25; Hunt 2000; Vihvelin 2000; and Widerker and McKenna 2003].

[2]The various (incompatibilism-motivated) objections presented by Widerker, Kane, van Inwagen, Ginet and O'Connor among others are well-known to anyone familiar with this literature (for references to these see [Fischer 1999; Hunt 2000; and Widerker & McKenna 2003]). There are other, less widely discussed objections to be found in [Cain 2003], who questions whether Frankfurt-style cases describe a genuine metaphysical possibility, and in [Vihvelin 2000]. I return to Vihvelin's objection below.

[3]I agree with Frankfurt's remark in that paper that it is not clear whether someone 'who accepts [the Principle] is thereby committed to believing that moral responsibility and determinism are incompatible' [Frankfurt 1969: 829].

[4]See [Fischer 1994: 131 – 59], and more recently [Fischer 2003]. See also [Hunt 2000] and [Pereboom 2003].

person is morally responsible for what he has done only if he could have done otherwise' [Frankfurt 1969: 829]. Let me say something about the interpretation of the Principle.

I take it that when the Principle is expressed by saying that a person is morally responsible for an action only if 'he could have done otherwise', what is meant is simply that the person is not morally responsible unless it was possible for him *not* to do what he did. That is, the Principle does not, in addition, require that the agent could have performed *some other action* instead. Or, to put the point differently, there are two ways in which someone who performs an action could do otherwise, or act differently. One is by doing something else; the other is by simply not performing the original action. I take the Principle to say that moral responsibility requires only the second kind of possibility: the possibility of not performing the original action.

It is, admittedly, possible to interpret the Principle more strongly and to take it to say that moral responsibility for an action requires that the agent should be capable of performing a different action. However, I cannot see any reason to interpret the Principle thus, nor do I think that this is how the Principle has been traditionally interpreted.[5]

Against this, it may be argued that the Principle surely requires more than the possibility of merely not performing the original action: it requires the possibility of *refraining* from performing the original action. And this, in turn, requires the possibility of deciding, forming the intention, or choosing not to perform the original action. And since deciding, choosing or forming an intention are mental actions, the argument goes, it follows that the 'could have done otherwise' clause of the Principle requires that the agent could have performed *some other action*—if only a mental act of deciding, choosing, etc., not to perform the original action.

I agree that the Principle requires that the agent should be able to *refrain* from performing the original action. But I see no reason to accept that the possibility of refraining requires the possibility of 'performing a mental act' of deciding not to act. For it seems plausible to argue that an agent who ϕ-s *had* the possibility of refraining from ϕ-ing if at t, when the agent ϕ-s, it was open to him not to ϕ, and it was up to him whether he ϕ-ed or not. And it is also plausible to argue that it was up to him whether he ϕ-ed or not if his ϕ-ing depended on whether he decides (or chooses, etc.) to ϕ. If these conditions are met, then an agent who ϕ-s at t could have refrained from ϕ-ing at t.

It may be objected that the possibility of refraining from ϕ-ing required by the Principle is the possibility of *intentionally* refraining from ϕ-ing. And, the objection continues, for someone's refraining from ϕ-ing to be intentional, his refraining must *be caused* by a mental act of deciding not to ϕ. So the Principle does after all require that the agent should be able to perform some other action. The objection is unconvincing, however, because it relies on the doctrine that everything we do or fail to do

[5]Indeed, even critics of the Principle such as Fischer would seem to agree. For he says that 'regulative control' requires 'the power to freely do some act A, and the power freely to do something else instead', with the following parenthesis: where 'doing something else' may be simply refraining from acting at all or 'doing nothing' [Fischer and Ravizza 1998: 31].

intentionally must be caused by a mental act of deciding (or choosing, etc.) to do it, or not to do it. But although the doctrine is very popular, it is at best problematic. To mention just two reasons that make it so: if deciding (or choosing, etc.) to ϕ is itself something we do intentionally, the view seems to generate a vicious regress. Besides, many times it is implausible to claim that my deciding (or choosing, etc.) to ϕ intentionally and my ϕ-ing intentionally must be two discrete happenings: my deciding or choosing to have another chocolate may simply consist in my having it when I could have refrained from doing so. And my deciding or choosing not to have it may consist simply in my not having it when faced with the possibility of having it. Now, I do not pretend that these brief remarks settle the debate on that doctrine but they are enough to show that the objection is not decisive, contrary to what those who embrace the doctrine might think.

So, on a plausible interpretation, which is the one I shall defend, the Principle says that an agent is not morally responsible for ϕ-ing at t, unless he could have refrained from ϕ-ing at t, i.e. unless, at t, it was up to him whether or not he ϕ-ed.

I now turn to examine Frankfurt's own alleged counterexample to the Principle.

III. Frankfurt's Thought-experiment

Both in the original and in subsequent papers, Frankfurt gives more than one version of his alleged counterexample. However, the general form of the received counterexample is as follows:

> Suppose someone—Black, let us say—wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones *decides to do, and that he does do*, what he wants him to do. Whatever Jones's initial preferences and inclinations, then Black will have his way. ... Now suppose that Black never has to show his hand because Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform.
>
> [Frankfurt 1969: 835–6]

Then, according to Frankfurt, since Jones acts 'for reasons of his own', he is morally responsible for performing the relevant action, despite 'the fact that he could not have done otherwise' [ibid. 836]. For in this case, Frankfurt claims,

> what action he [Jones] performs *is not up to him*. Of course it is in a way up to him whether he acts on his own or as a result of Black's intervention. That depends upon what action he himself is inclined to perform. But whether he

finally acts on his own or as a result of Black's intervention, he *performs the same action*.

[loc. cit. My italics.]

So, he concludes, this is a counterexample to the Principle. Is Frankfurt right?

IV. The Conditions for Counterexamples to the Principle

Given the interpretation of the Principle outlined above, a Frankfurt-type counterexample to the Principle must be a case where:

(1) the agent performs an action for which he is morally responsible; and

(2) the agent *could not* have avoided performing that action.

There are two points I want to emphasize about these conditions, in particular about condition (2).

The first point is that both what the agent is morally responsible for and what he cannot avoid doing must be *performing an action*—indeed the same action in both cases.[6] And this means that for a case to be a counterexample to the Principle, condition (2) requires that what Jones would do in the counterfactual case, i.e., as a result of Black's intervention, should be rightly describable as 'performing an action'. That is, it would not be enough if in the counterfactual case Black simply ensured that Jones's body moved in this or that way (e.g. the way it would move were Jones to perform the relevant action) because if he merely ensured that, Black would not have ensured that Jones *acted*.

This is so because it is not enough for someone to have performed an action, or indeed to have moved his limbs, that his limbs move, even if they move in exactly the same way as they move whenever he performs an action of that kind. Following Hornsby [1980] we can use the subscripts '$_T$' and '$_I$' to indicate whether a phrase should be understood as corresponding to the transitive or intransitive form of a verb. So if I melt a piece of chocolate, we can distinguish between what I do, melt the chocolate; and what happens to the chocolate, that it melts. And we can use the nominals 'the melting$_T$ of the chocolate' to refer to the first; and 'the melting$_I$ of the chocolate' to refer to the second. Similarly we can use the nominal 'bodily movement$_T$' to refer to my action of moving my body, and the nominal 'bodily movement$_I$' to refer to what happens to my body, that it moves. So, it is not enough for someone to have performed an action that some bodily movements$_I$ occur, even if those are the exact movements$_I$ that occur when he performs an action: what is needed is that the agent should act—i.e. that there should be some bodily movements$_T$. Consider the following examples. Suppose that an earthquake makes Jim's limbs move in the way in which they move whenever he dances the samba. The fact that his limbs moved so is not

[6]But unlike Ginet [1996] and others, I don't mean he must perform the same 'particular' action. In the counterfactual case the agent would have to perform an action of the same kind as that for which he is held responsible in the actual case.

enough to conclude that Jim danced the samba, nor to conclude that Jim moved his limbs, for the earthquake caused Jim's limbs to move, but it did not cause Jim to move his limbs. So, all that can be concluded is that the earthquake caused Jim's limbs to move in the way they move when he dances the samba, i.e. that the earthquake caused some bodily movements$_I$. Likewise, if C presses A's trigger finger while A is holding a loaded gun and B gets shot as a result, it would be wrong to conclude that C caused A to shoot B. What C would have done is cause *A's finger to move* in the way in which it would have moved had A shot B (i.e., it would cause a finger movement$_I$). But that does not amount to causing *A* to shoot B. In fact, in such a case, it would be C that would have shot B. In neither of these cases, then, could it be said that the agent (Jim and A respectively) had *acted*.

This is not always made clear. Consider, for example, what Frankfurt says about alleged counterexamples in a later paper:

> The distinctively potent element of this sort of counterexample to PAP is a certain kind of overdetermination ... The arrangement ensures that a certain effect will be brought about by one or the other of the two causal factors, but not by both together. Thus the backup factor may contribute nothing whatever to bringing about the effect whose occurrence it guarantees.
>
> [Frankfurt 1982] quoted in [Hunt 2000: 224, n.12]

Frankfurt's way of putting the point is ambiguous, however, on whether the 'determination' of the case concerns an *action* (A's killing B) or its *result* (B's death).[7] But a counterexample to the principle requires that the determined effect should be an *action* that the agent performs, and not simply its result. The reason for this is that the Principle does not say that an agent is morally responsible for the result of his action only if he could have prevented that result, but rather that he is morally responsible for his action, and its result, only if he could have acted otherwise, i.e. if he could have refrained from performing that action.

The second observation about condition (2) is that it requires that the agent *cannot avoid* performing the action. This is also made explicit by Frankfurt when spelling out the kind of effective steps Black could take to ensure the right outcome. Again Frankfurt offers more than one suggestion,[8] but the most promising one, and the one that has captured the imagination of philosophers, is that Black would manipulate

> the minute processes of Jones's brain and nervous system in some direct way, so that causal forces running in and out of his synapses and along the poor man's nerves *determine* that he chooses to act and that he does act in the one way and not in any other.
>
> [Frankfurt 1969: 835–6. My italics]

[7] I use the word 'result' here in von Wright's sense in which the connection between an action and its result is logical, that is, an action is of such and such a kind (e.g. a killing) if and only if its *result* is of the corresponding kind (a death) [von Wright 1963: 39].

[8] The others involve making an extremely severe threat or hypnotizing Jones, both of which, for different reasons, would make the claim that in the counterfactual case the agent couldn't have avoided performing the action highly implausible. More on this below.

Thus, for condition (2) to be met, Black would have to ensure not only that Jones performs an action but also that Jones *could not have avoided* performing the action. Let me spell out what this implies.

First, it implies that it would not be enough if the background conditions were such that Black would merely cause Jones to perform the action, for this would only sanction the conclusion that Jones *would have* performed the action either way, not that he could not have avoided performing it. This second conclusion follows only if Black's intervention was irresistible: i.e. only if it was not up to Jones whether he performs the action or not.

One might object that, surely, if Jones will perform the action whether he does it for reasons of his own or as a result of Black's intervention, then Jones cannot avoid performing the action. But, unless a premise is added to the effect that Black's intervention is irresistible, this reasoning involves a modal fallacy, for, from a premise that something will happen, it is not legitimate to infer that nothing else could have happened. Consider the following: 'He went go to London by train. And, had he not gone by train, he would have gone by coach.' From these premises we can conclude that, one way or another, he would have gone to London. But we cannot conclude that he could not have avoided going to London. Likewise, 'Jones ϕ-ed for reasons of his own, but had he not done so, he would have ϕ-ed as a result of Black's intervention', does not imply that Jones could not have avoided ϕ-ing—unless Black's intervention was irresistible.[9]

So, for a Frankfurt-style case to succeed as a counterexample, it is necessary that in the counterfactual case Black would *determine* that Jones performs the relevant action; that is, he must cause Jones to perform an action in such a way that Jones cannot avoid performing it. Only if those were the background conditions of Jones's action would it be true that it was not up to Jones which action he performed, and hence only then would this case satisfy condition (2).

It should be noted here that it is not legitimate for a Frankfurt-style case simply to *stipulate* that, in the counterfactual case, the agent would be caused to perform an action that he cannot avoid performing (i.e. to stipulate that condition (2) would be met). Rather, any example needs to tell a compelling story that makes the suggestion plausible without begging the issues at hand.[10] For the idea that this possibility is conceptually

[9]Or so most people would think. But Vihvelin [2000] argues that even this would not be enough and that the thought that it would also rests on a modal fallacy. Her overall argument against Frankfurt-style cases is complex and subtle, depending as it does on a distinction between conditional and counterfactual intervention, and I cannot do justice to it here. What I want to highlight here is her claim that, even if in the counterfactual case Black could cause Jones to act so that the latter cannot avoid acting, this is not enough to show that, in the *actual* case, Jones could not have done otherwise and hence not enough for a Frankfurt-style case to succeed. I am inclined to think that Vihvelin's objection is right and hence devastating. But I shall not rely on it here, and shall instead argue that, even if she is wrong, Frankfurt-style cases still fail.

[10]So a plausible story needs to be told to the effect that the agent 'could not have done otherwise'—i.e. a story that is not, e.g., conceptually problematic, question-begging, or dependent on highly controversial claims. Thus, for example, D. Hunt [2000], claims to offer counterexamples to the Principle that do not rely on any Frankfurt-style counterfactual controller. However, his alleged counterexamples all suffer from one or several of those defects. One of his examples crucially involves the possibility of backwards causation—but, if that is not conceptually problematic, what is? Another depends on the view that the existence of 'infallible true beliefs' about what someone will do implies that the person could not have done otherwise—but this view is (and, for centuries, versions of it involving knowledge have been) highly controversial. Hunt's third example also involves problematic assumptions (e.g. about the relation between mind and brain) but, in any case, it fails as a counterexample to the Principle (see n. 12).

unproblematic is by no means self-evident: as I shall argue in Section VI, the concept of what performing an action is makes the cogency of this idea at best highly dubious.

So I shall show below why Frankfurt-style cases cannot meet condition (2), but before doing that I need to make an argumentative aside. I have described what would count as a counterexample to the Principle as Frankfurt and many of his followers conceived of them. However, as I noted in the Introduction, it has been suggested more recently that the Principle could be falsified by a different kind of case, and I ought to say something about that before proceeding with my arguments about Frankfurt-style cases.

A. 'Fischer-style' Cases and the Principle

I said above that a Frankfurt-style counterexample to the Principle would be a case where the agent either ϕ-s for reasons of his own, or is inescapably caused to ϕ; i.e., to use Fischer's terminology, cases where 'the agent is restricted to pathways with the same contents' [Fischer 2003: 241]. However, according to Fischer, a counterexample to the Principle need not be a Frankfurt-style case but could instead be a 'Fischer-type' case. The latter are cases where the agent '*lacks access* to pathways along which there are *relevantly different* contents' [ibid.]; i.e. Fischer's are cases where the agent either ϕ-s, or can do *nothing* else.

Now, my aim in this paper is to show that Frankfurt-style cases *must* fail as counterexamples because they are conceptually flawed. Fischer-style cases do not have the same flaw but I think it is nonetheless possible to show that they also fail as counterexamples to the Principle. This is Fischer's example:

> In this kind of variant on the Frankfurt-type scenario, if it becomes clear to the counterfactual intervener—he is an excellent judge of such things—that the agent is going to decide to do something else, he will use his machine to destroy the agent's brain and thus kill him instantly!
>
> [Fischer 2003: 242]

According to Fischer this is a counterexample to the Principle because it is a case where the agent would be morally responsible for what he actually does, since he does it for his own reasons, even though he could not have acted otherwise because he could not have performed any other action.

The claim that this kind of case is a counterexample to the Principle depends on interpreting the 'could have acted otherwise' clause to mean 'could have *performed a different action*'. But, as we saw above, the clause need not be so interpreted because there are two ways in which someone who performs an action could do otherwise, or act differently. One is by doing something else; the other is by simply not performing the original action (and as I point out in note 5, this is something that, given what Fischer says elsewhere, he ought to agree with). Now, although in Fischer's example it is not possible for the agent to perform a different action—Black

is there to prevent that—it is possible for him *not* to perform the original action. So he could have acted otherwise.[11]

Those familiar with the literature might think that this response to Fischer is open to his 'flicker of freedom' objection, namely that not *any* alternative will do as a defence of the Principle. As he puts it, some alternatives represent a mere 'flicker' of freedom and they do not have 'sufficient ''oomph'' to ground moral responsibility' [Fischer 2003: 242]. And, one might think, my response is open to this objection because it relies on the fact that the agent had *an* alternative, but not one with 'sufficient ''oomph''': for my response points out that it was possible for the agent *not* to perform the original action but that is surely not a sufficiently robust alternative.

But this interpretation would involve a misunderstanding of my response. For my response is that, in a Fischer-style case, the agent who ϕ-s *could have refrained* from ϕ-ing, e.g. could have refrained from shooting Smith. It is true that Black's presence in the background prevents the possibility of a mental act of deciding, or choosing, not to shoot Smith. But, as we saw above, we have no reason to accept that the possibility of such a mental act is necessary for the possibility of refraining to have been present. What is necessary for the possibility of refraining is that at the moment when the agent ϕ-s, the agent could have not ϕ-ed, and that it was up to him whether he ϕ-ed or not; for instance, that at the moment at which Jones shoots Smith, Jones could have not shot Smith, and that it was up to him whether he shot Smith or not. And in Fischer-style cases at the moment when Jones shoots Smith, Jones could not have shot him, and *it is up to Jones* whether he shoots Smith or not: if he decides to shoot he shoots him, and if he does not decide to shoot, he does not shoot him. Black's presence in the background does not affect this fact. And that is all that is required for Jones to have been able to refrain from acting and hence to have been able to act otherwise.[12]

Thus Fischer-type cases are cases where the agent has access to pathways with 'relevantly different contents'—for shooting Smith is as different from not shooting Smith as the contents of two pathways can be; and it is up to the agent whether he does perform the relevant action or not, for whether he does depends on whether he decides to shoot Smith, or whether he refrains from deciding to shoot him. Because of this, Fischer-type cases meet the

[11]Pereboom [2003] contains an excellent discussion of other objections to Fischer's example.

[12]The same kind of response can be made to Hunt's third case. He describes a case where Jones murders Smith for his own reasons but where, Hunt claims, he could not have done otherwise because all neural pathways in Jones's brain are blocked *except* for the pathway that needs to be open for Jones to murder Smith (i.e. the only route which is open by sheer chance is 'precisely the route the man's thoughts would be following anyway' [Hunt 2000: 218]. But Hunt's claim is unconvincing because, even if, given the state of his brain, Jones could perform no *action* other than murdering Smith, Hunt has not given any reason why Jones could not, at any point in the series, simply refrain from murdering Smith. To be sure, if Jones's refraining required the occurrence of some neural event other than those in the open route, that might be ruled out. But why should we think it does? Why is it not possible that his refraining should simply require the open neural series' coming to a stop? And, since Hunt gives no conceptual or empirical grounds why this should not be possible, for him to *stipulate* that Jones could not refrain is simply for him to stipulate that this is a counterexample to the Principle; but this is not to tell a plausible story that could persuade us that it might be. The same is true of Pereboom's attempt to reinforce a similar counterexample by stipulating that 'it is causally determined that [the agent] remain a living agent, and if she remains a living agent, some neural pathway has to be used' [Pereboom 2001: 16]. Again, it is not clear that this rules out the possibility that the only open neural pathway should come to a stop and that that possibility is all that is required for the agent to be able to refrain from acting.

'could have done otherwise' condition and are not, therefore, counter-examples to the Principle.

I shall now return to Frankfurt-style cases in order to examine whether *they* are counterexamples to the Principle, and in particular whether they can meet the two conditions stated above: that the agent does something he is morally responsible for, and that he cannot avoid doing that thing.

V. Mental Acts and Brain Events

As I mentioned above, in the original Frankfurt example we are told that Black would manipulate

> the minute processes of Jones's brain and nervous system in some direct way, so that causal forces ... *determine* that he chooses to act and that he does act in the one way and not in any other.
>
> [1969: 835–6. My italics]

In other words, we are told that Black would *determine* that Jones chooses or decides to act.[13]

Now, if, by manipulating Jones's brain, Black could determine that Jones decides to act, then we would seem to have a counterexample to the Principle. For deciding to do something can itself be regarded as a mental act for which we may be held morally responsible. And therefore this would seem to be a case where an agent does something for which he is morally responsible, namely, he decides to shoot Smith, but where he could not have refrained from doing it, that is, he could not have refrained from deciding to shoot Smith.

However, the suggestion that Black *could* cause Jones to decide to act by manipulating his brain requires closer examination.

First, one may question whether it is *conceptually legitimate* for a thought-experiment to stipulate that manipulating someone's nervous system can amount to causing him to make a decision. The suggestion may seem plausible if one thinks that the connection between the occurrence of brain events and that of mental acts is such that causing a brain event might amount to causing a decision. But this line of thought is highly problematic. For suppose that Black can cause certain events to occur in Jones's brain: what grounds are there for calling the brain events that Black would thus cause a *decision* of Jones's?

The grounds could not be, for instance, that Jones himself would assert, outwardly or *in foro interno*, that he had made a decision. For Jones might be wrong, and his assertion might simply be a further effect of the brain manipulation. Indeed, his rehearsing those words might not amount to an assertion at all.

[13]Frankfurt sometimes talks about 'Jones's choice' and sometimes about 'Jones's decision'. Most people, whether critics or not, talk about decisions and I'll do the same, but I believe nothing depends on this point. Also, in Frankfurt's paper the protagonist of the counterexample is Jones_4 but for ease of exposition I shall refer to him simply as 'Jones'.

Nor can the reason be that Black would have succeeded in replicating the neural event, or sequence of neural events, which would have occurred had Jones made a decision on his own. For suppose that he succeeded in doing that. Why would making this sequence happen amount to causing *a decision*? Just as the occurrence of a sequence of movements$_I$ of Jones's body is not sufficient for his having performed an action, the occurrence of a particular sequence of neural events is not sufficient for his having made a decision—*even if*, had Jones made a decision, that same sequence would have happened; and even if, whenever Jones makes a decision to perform an action of that kind, a sequence of neural events of the same kind occurs.[14]

Thus there seem to be no grounds for thinking that it is legitimate to stipulate, as Frankfurt and others do, that what Black would cause by manipulating Jones's brain would be a decision of Jones's. Indeed, there are compelling reasons for denying that it is legitimate, because none of the criteria for someone to have decided something would apply to Jones in the counterfactual case. For one thing, the 'decision' would not be the outcome of Jones's practical reasoning, or of his emotional response to a situation. But it is in such contexts, i.e. contexts of deliberation, of appraisal and reaction to a situation, etc., that the concept of a decision has application.

Moreover, a crucial (though defeasible) criterion for whether someone has decided to perform an action is whether they do, or try to, act according to the alleged decision. But here we cannot use the fact that Jones goes on to do something (say shoot Smith) as grounds for saying that he had indeed decided to do so. This is because, here, the only reason we might have for describing the relevant movements$_I$ of Jones's body as connected to his *doing* something, i.e. his shooting Smith, is that we are told that these movements$_I$ are the result of a *decision* to do so. As will become clear below, the concept of a decision is introduced in the example precisely in order to undercut the objection that all Black would do is cause Jones's body to move, and not, as a counterexample requires, cause him to move his body. But this means that one cannot help oneself to the consideration that Jones would *act* according to his decision in order to bolster the claim that what Black had caused was indeed a decision.

It seems, then, that we have reason to deny that it is cogent to say that Black could cause a decision in that way. By manipulating his brain, Black can *prevent* Jones from deciding *not* to perform the action (e.g. Black could sedate Jones, or kill him). But preventing A from deciding not to kill B is not the same as causing A to decide to kill B. The first is conceptually unproblematic (if fanciful science-fiction as described in Frankfurt-style cases); the second, however, is conceptually problematic and the thought-experiment cannot simply stipulate that, by manipulating Jones's brain, Black could cause Jones *to decide*.[15]

[14]And even if one thinks that the occurrence of a decision is *identical* to the occurrence of a neural event, still the occurrence of a neural event of the relevant kind would not be sufficient for the occurrence of a decision. The point can be modified to fit whatever relation is thought to obtain between the neural event and the decision, e.g. that the former is the 'realizer' of the latter, etc.

[15]These considerations undermine Mele and Robb's claim [1998] to have rescued Frankfurt-style cases from objections that say that it is not legitimate to predict a decision, by showing that a decision could be randomly caused by a device without the need to predict anything. For their arguments assume that it is conceptually legitimate to stipulate that decisions can be caused by causing brain events.

The defender of Frankfurt-style cases may suggest at this point that Black could cause Jones to decide in some other way: by persuasion, threats, etc. And if Black could do that, then the case would constitute a counterexample to the Principle. I examine this suggestion in the following section.

VI. Causing Unavoidable Actions?

The suggestion under consideration is that Black might have means other than causing brain events by which to ensure with the required necessity that Jones acts. I shall examine this suggestion, leaving aside the issue whether the action thus caused would be a mental act (of deciding, or choosing), or an action involving changes in Jones's body, for the arguments that follow apply to both.

The idea that in the counterfactual case Black would cause Jones to act with the necessity required by condition 2 above, presupposes that it is conceptually legitimate for a thought-experiment to stipulate that Black could cause Jones to act in such a way that condition (2) is met. But it is far from clear that it is, because it is far from clear that the idea that a person can be caused to perform an action in such a way that he could not have avoided performing it but which is nonetheless *his* action is cogent. As Geach puts it:

> If some action on a man's part is wholly determined by ... events and circumstances in the world over which the 'agent' had no control, then it is quite inappropriate to call him an agent.[16]

(And, as we shall see below, Frankfurt's conception of what it is to act, as articulated e.g. in [Frankfurt 1978], supports this point.)

Now, it is certainly possible to cause people to (decide to) perform actions—for example, through entreaties, promises, requests, reasoning, incitements, threats, etc. However, when actions are thus 'caused', they are not actions their agents cannot avoid performing, and therefore this is not the kind of causing required for Frankfurt-style cases to succeed. This is because these ways of causing someone to act work by *persuading* the agent to do something, and the need for persuasion arises from the fact that what he does is up to the agent: 'an offer you cannot refuse' is, precisely, a joke. Hence, in as much as the action is caused *by those means*, it is open to the agent to refrain from acting—though admittedly sometimes at great cost.

But surely, it may be objected, there may be bribes so tempting, or better, threats so terrible, that it may be impossible to resist. So faced with such a threat, the agent may not be able to act otherwise. But the idea that there are *irresistible* threats is mere hyperbole. A threat works by making non-compliance, i.e. the *alternative* course of action, highly unpalatable to the agent—not by eliminating its possibility.

[16][Geach 2000: 80]. See Aristotle [1984: *Metaphysics*, 1048a6ff.], Aquinas [1960 – 73: 1a, 2ae, 49, 4], and Reid, who says that 'power to produce any effect implies power not to produce it' [Reid 1969: 1.v.35]—and he adds that 'otherwise it is not power but necessity' [ibid.]. The idea is also developed in [Kenny 1975: ch. 7].

That threats can be resisted becomes evident if one thinks of examples where the threat is indeed terrible, but the action required is equally terrible. So someone threatens to torture your child unless you torture another child. This is an awful predicament *not* because you have no choice, but because you *do*; this is not an irresistible threat but rather an appalling choice.

It is true that sometimes people are not blameworthy for things they do under threat. But that is not because they couldn't have done otherwise; rather it is because there are some threats one oughtn't to resist, or at least where it is not true that one ought to resist them.[17]

So people can be compelled to do things through threats, bribes, etc., in which case they will do what they do more or less unwillingly, even with repugnance. And while this sort of compulsion is of course quite relevant to assessments of moral responsibility, it does not imply that the agent could not have avoided performing the action.

Thus, if the way Black would cause Jones to act in the counterfactual case were through threats, bribes, etc., condition (2) would fail because that would not be a case where Jones did something that he could not avoid doing.

In a 1978 paper Frankfurt says that what is required to accept the conceptual possibility of a counterexample to the Principle is to accept that 'it is possible that an action should be caused by alien forces alone' [Frankfurt 1978: 50]. But, as I just argued, that understates the case: what is required is that an action should be caused by alien forces *in such a way* that the agent cannot avoid performing the action. And the question is whether this suggestion is cogent.

One might think that it is surely cogent to suppose that someone can be caused to act in such a way that he could not have avoided performing the action. For surely agents sometimes act but cannot avoid doing what they do, for instance, in the case of obsessive-compulsive behaviour, or of some actions performed during psychotic episodes, or under the grip of irresistible desires. And if that is so, all we need to imagine is that, in the actual case, the agent performs an action of that kind for reasons of his own, while in the counterfactual case, something or someone would bring it about that the agent becomes an obsessive-compulsive, or becomes psychotic, or is in the grip of an 'irresistible desire', and performs one of those unavoidable actions. If so, the agent would have been caused to perform an unavoidable action.

But it is not clear that the examples mentioned above are really actions that are 'unavoidable' in the required sense. In the case of obsessive-compulsive behaviour, such as compulsive washing, counting, etc., the truth seems to be that, for any particular occasion, the agent can avoid acting. The compulsion, such as it is, lies not so much in each action but in a pattern of irrational, because unnecessary or harmful, repetition.[18] For it should be

[17]Very often people don't resist threats, but it doesn't follow that they couldn't have. People often say: if I'd been stronger (more ruthless, more generous, less ambitious) I could have resisted that threat. But, since there is no independent measure of whether you were strong enough to resist the threat, the notion of 'strong enough' does not really refer to an antecedent enabling condition. Your past actions might give us an indication of what you are likely to do faced with that kind of threat but not of what you *can* do—people can always surprise us.

[18]The same is true of the actions of the addict: the inclination to smoke *this* cigarette, or to inject *this* dose, is itself conquerable. What makes kicking a habit so hard—though possible—is, first, the need for suitable motivation, and second, the difficulty in sustaining the motivation long enough to conquer the addiction.

noted that these conditions are typically treated with cognitive-behavioural therapies—therapies that work by increasing awareness of the behaviour and its psychological origins, replacing the behaviour with other activities, etc., all of which is aimed at removing the habit of repetition. But those therapies could not even begin to be deployed if the actions that form the pattern were unavoidable in the sense that the Principle requires.

Concerning psychotic episodes, there is even less reason to imagine that acts committed during such episodes are acts that the agent could not avoid performing. It is true that agents are often held not to be morally responsible for those acts but that is not because it is thought that they could not avoid performing them. The reason is, rather, that agents in those conditions lack what H.L.A. Hart calls 'capability responsibility', i.e. the cognitive capacities required for moral responsibility [Hart 1968]. Psychotic episodes bring with them a disconnection (in perception, thought, etc.) with reality, and that impairs the agent's capacity to make judgements about the character of their behaviour, about right and wrong, and so on. So although these agents are often excused from moral responsibility, this is not because they cannot avoid doing what they do but because, in the moral sense, they do not *know* what they are doing.

Sometimes the term 'psycho' is used for agents who engage in random acts of extreme or irrational violence. In those cases, however, it is often not even true that the agent is held not to be morally responsible. Such agents are sometimes said to be 'amoral', not because they cannot tell moral right from wrong but because they don't care about the distinction. So again, there is no reason to think that these are cases where the relevant agents cannot avoid doing what they do. Rather the contrary: they could refrain but they lack any motivation to do so.

What about 'irresistible desires'? Mele writes:

> Irresistible desires are mentioned with unsurprising frequency in discussions of free agency and moral responsibility. Actions motivated by such desires are standardly viewed as compelled, hence unfree. Agents in the grip of irresistible desires are often plausibly exempted from moral blame for intentional deeds in which the desires issue.
>
> [Mele 1992: 86]

But, despite the alacrity with which the term is used in those discussions, the concept of irresistible desire calls for some critical examination.

We do, outside of philosophy, talk about 'irresistible desires', but when we do, we do not mean that they are irresistible in the sense required by the Principle. For we often say things like 'I felt an irresistible desire to ϕ' e.g., to slap him, to kiss her, to smash the vase, etc., but then add 'but I managed to overcome it' without any sense of real contradiction. The reason for this is that, in ordinary use, the term 'irresistible desire' is merely a *façon de parler*: we often resist 'irresistible' desires, just as we tolerate 'intolerable' situations, suffer 'insufferable' people, or believe 'unbelievable' claims. And one is not normally exempted from moral blame for the intentional deeds in which such 'irresistible' desires issue: a judge is unlikely to be impressed by

the excuse that a man shot someone because he felt an irresistible desire to do so. The judgement may be affected by whether the person was aware of what he was doing and its implications, by whether there was premeditation, or severe provocation, by whether the agent 'had lost his mind', etc. but, as we saw above, this is not because the judge will then believe that he could not have resisted the desire to shoot the man but rather because the agent would have lacked, at the time or permanently, 'capability responsibility'.

So, at least in ordinary talk, an 'irresistible' desire is not a desire that one could not resist but rather one that was keenly felt and took a great deal of effort to overcome—or not. Indeed the idea that desires are literally irresistible, rather than merely not resisted, is problematic. Because for a desire to be literally irresistible by someone is for that person to lack the *capacity* to resist it. But there does not seem to be any criterion for whether a person had the capacity to resist a desire that is independent of whether they resisted it. And yet, the fact that a person didn't resist a desire does not license the conclusion that the person lacked the capacity, i.e. that the desire was irresistible to him, rather than resistible but not resisted. But without such a criterion, what does it mean to say that a desire was irresistible, as opposed to resistible-but-not-resisted? What is the difference between describing a desire as one or as the other?

In fact, it has been plausibly argued that the concept of desire involves the idea of an inclination that *can* be resisted. For instance, Kenny says that

> wants can be invoked to explain an agent's action only when it is in the agent's power to act in a manner other than that which amounts to a fulfilment of the want. This is an essential element in the concept of 'want' and in the procedure of explaining actions in terms of wants, whatever form the want is in question, whether purpose or intention, whether volition or desire.
>
> [Kenny 1989: 47]

And he adds, 'wants are attributed to people on the basis of what they do when it is open to them to do otherwise' [ibid. 48]. Perhaps there are inclinations to do something that cannot literally be resisted, but then the inclination is not a 'desire', and the resulting behaviour often hardly qualifies as an action of which one can be said to be the agent, instead of a reaction, a reflex movement, or even a bodily function (more on this below). And the reason why we call such behaviour 'reactions' or 'reflex movements' is precisely that they are beyond the agent's control, and result from urges or impulses to do something (breathe, blink, giggle, kick, etc.) that cannot be resisted.

Thus, none of the suggestions examined seem really to be examples of actions of the right kind and that are unavoidable in the sense required by the Principle. And this is no coincidence, because the concept of what it is for someone to act, i.e. for something to be an action of which one is the *agent*, makes the idea that it is possible to cause the relevant actions so that they are *unavoidable* problematic. It seems that for what someone does to be *his action*, for him to be its *agent*, the person must have a certain degree of control over it: at least, he should be capable of refraining from doing what

he does. But a counterexample to the Principle requires that the 'agent' have no control over what he would do in the counterfactual case, even though what he would do would be an action of his—and an action of the kind for which one could be held to be morally responsible. And these are inconsistent requirements.

It is true that it is possible to cause people to do things that perhaps they cannot avoid doing: say giggling (by tickling), blinking (by bringing an object close to one's eyes), vomiting (in a variety of ways), and a whole raft of more or less reflex movements that can only be suppressed, if at all, with tremendous effort and if one is forewarned. However, the things that can thus be caused are mostly on the borderline between what one does and what happens to one. And when they are things that one does, they are either not really *one's* actions (rather they are actions of parts of one's body), or they are actions but not the kinds of action for which one might be held to be morally responsible, and hence these doings would not serve to construct counterexamples to the Principle.

It may be tempting to think that surely they would serve, for we can imagine cases where, say, giggling or blinking would constitute a pre-arranged signal, and hence something for which one could be morally responsible, and then add that the counterfactual intervener could cause the agent to giggle or blink in such a way that the agent could not stop himself from doing those things. So here we seem to have cases where the person does something for which he is held morally responsible and which he could not have avoided doing.

But the appearance is deceptive, and these cases are not counterexamples because the action the agent is morally responsible for is *not* the same action (i.e. not an action of the same kind) as that which he could not avoid performing. So one might give a signal by blinking, in which case one may be morally responsible for *giving the signal*—which one did by blinking. However, if in the counterfactual case one was 'unavoidably' caused to blink, one would not *thereby* have given a signal, in the same way in which although one can make a bid at an auction by making a nod, not any nod at an auction will constitute a bid: the agent must intend the nod as a bid. Likewise, for a blink to be a signal, the agent must intend the blink as a signal. So, if in the counterfactual case the agent was caused to blink but did not intend it as a signal, he would not have given a signal, and hence this would not be a case where he was morally responsible for something (giving a signal) that he could not avoid doing – because, although he could not have avoided blinking, he *could have* avoided giving a signal, which is the action for which he is morally responsible. And since, as we saw in section V, there is no reason to think that the agent could not refrain from intending the blink as a signal, it follows that the agent could have avoided giving a signal.

VII. Agency and Actions

In the 1978 paper mentioned above, Frankfurt himself provides a characterization of what it is for someone to act that strongly suggests

that he ought to deny that one could be caused to perform an unavoidable action. For Frankfurt's view is that someone performs an action (involving moving one's body) *only if* his movements are under his guidance.

> What is not merely pertinent but decisive, indeed, is to consider whether or not the movements as they occur are *under the person's guidance*. It is this that determines whether he is performing an action. ... What counts is whether he was prepared to intervene if necessary, and that *he was in a position to do so more or less effectively*. ... The assertion that someone has performed an action entails that his movements occurred under his guidance.
>
> [Frankfurt 1978: 45–50]

And he goes on to say that someone's movements are under his guidance *only if* the person is in a position to intervene and change things if 'the accomplishment of its course were to be jeopardized' [ibid.].

Frankfurt seems to be using the word 'movements' here in a general sense: in the sense in which, e.g. the police ask people about their movements on the day of the crime; so he is using it in a sense that includes but is not restricted to bodily movements$_T$. Be that as it may, this characterization of what it is for someone to act presents a problem for his alleged counterexample because, as we saw, condition (2) requires that the agent in a genuine counterexample should be wholly unable to intervene, and therefore, by Frankfurt's own criterion, it requires that the person should not be able to perform an action.[19]

Now, one need not accept Frankfurt's picture of agency to agree that, for something that happens to count as *one's* action, one must have *some* kind of control over it. But, since in the counterfactual case Black must have total control over what happens, Jones cannot have any control over it, and hence what happens cannot count as an action of Jones's, as something of which he is the agent—at most it might count as movements$_I$ of his limbs or as events in his brain.

In fact many people might think that Frankfurt's conditions for agency are too weak. They might think that for there to be an action of yours it is not enough that you are in a position to intervene and adjust or alter, e.g., the course of the movements$_I$ of your body. Rather, they might think, *you* must cause those movements$_I$. That is, if we consider the distinction between, say, raising your arm and your arm's rising, it might be thought that, in order for you to perform the action of raising your arm it is not enough that, as your arm rises, you are in a position to intervene and change its course if necessary; rather it might be thought that what is required is that *you* raise your arm—in other words, if to raise your arm is to make your arm rise, it must be you who makes it rise and not something or someone else. Consider something else Frankfurt says in the same paper:

[19]And it would not help to suggest that, in our example, Black's manipulation of Jones's brain would ensure that Jones's mental state at the time would be such that Jones *would not* want to intervene. It would not help because what matters is whether Jones would be in a position to intervene, not whether he would be inclined to. And, while for what Jones does to be an action the answer must be 'yes', for the counterexample to work, the answer must be 'no'.

> [T]he contrast between actions and mere happenings can readily be discerned elsewhere than in the lives of people. ... Consider the difference between what goes on when a spider moves its legs in making its way along the ground, and what goes on when its legs move in similar patterns and with similar effect because they are manipulated by a boy who has managed to tie strings to them. In the first case the movements are ... attributable to the spider, who makes them. In the second case the same movements occur but they are not made by the spider to whom they merely happen.
>
> [Frankfurt 1978: 51]

One might take issue with Frankfurt on the wording of his point, in particular with his claim that 'in the second case the same movements occur'. In the first case, it is $movements_T$ of its legs that are attributable to the spider, while in the second case what happens to the spider is that some $movements_I$ of its legs occur. In other words, and as Frankfurt acknowledges, there is something that happens in the first case, namely that the spider moves its legs, that does not happen in the second; that is, what happens in the first case that does not happen in the second is that there is an action of the spider's.

Similarly, if, by whatever means, Black caused certain brain events or certain $movements_I$ of Jones's body to occur, Black would not have thereby caused Jones to act. If, on the other hand, what Black causes is that Jones *acts*, then the latter would still have to have some control over this action, at least, the possibility of refraining from doing what he does, for otherwise what he does would not count as an action of *his*. So it would not be true that Jones could not have refrained from doing what he did.[20]

Thus, whether we adopt a more stringent criterion for acting or whether we accept Frankfurt's more minimalist account, the point is that in the counterfactual case, if Black causes something, say some brain event or some $movement_I$ of Jones's body, this would not amount to causing Jones to perform an action (not even to causing him to move his limbs—as the earlier example of Jim, the earthquake and the samba showed), and *a fortiori* it would not amount to causing Jones to perform an action that *he cannot avoid performing*.

Frankfurt-style cases are, therefore, faced with an insuperable difficulty. If, to give plausibility to the thought that Jones would indeed perform an action, they stipulate that Black would cause Jones to *act*, then we must conclude that Jones must have been able to refrain from performing that action, for otherwise that would not be an action of which Jones was the

[20]This is the kind of response one ought to give to anti-Principle strategies where, in the counterfactual case, Black would take control over Jones so that the latter would do what Black wanted him to, and *could not help* doing so. For first, the idea that someone could be controlled in this way is the stuff of Gothic novels or science-fiction. An excellent example of the genre is the film 'The Cabinet of Dr Caligari', where we are told that the latter has 'completely enslaved a somnambulist (Cesare) to his will, and compelled him to carry out his fantastic schemes', namely a series of murders. However, it is worth noting that in the story, the somnambulist turns out, after all, to be capable of avoiding doing the things that he is commanded to do. For, when it comes to it, he refrains from murdering the beautiful heroine. (Something similar is true of hypnosis: we can confidently say that people who are successfully hypnotized do what they are told, but this does not mean that we can conclude that they could not avoid doing it.) And this brings us to the second, related, point: the fact that the somnambulist must retain this power is what distinguishes him as an agent from a mere puppet or automaton. And this gives us a clue to an important conceptual point, namely that, if the fantasy of total control over someone became a reality, then it is not at all clear that the subject of such control would still remain an *agent*, or that what they 'did' would amount to *performing actions*.

agent. If, on the other hand, in order to ensure the required necessity, they stipulate that Black is in total control and causes something such that Jones is powerless to intervene, then Black might cause bodily movements$_I$, or brain events to occur, but then the claim that what would thus have been caused is that Jones *acted* collapses. Either way, condition (2) is not met. Frankfurt-style cases cannot succeed because, once we examine the concepts relevant to the case, it becomes clear that a counterexample to the Principle requires inconsistent conditions.

I conclude, then, that whatever Black can bring about, he cannot, because it is not a conceptual possibility, cause Jones to perform an action of which he is the agent, *and* at the same time make it the case that Jones could not have avoided performing that action. But that is what is required for conditions (1) and (2) to be met, i.e., for a Frankfurt-style case to be a counterexample to the Principle.

VIII. Conclusion

I have argued that Frankfurt-style cases must fail as counterexamples to the Principle because they require inconsistent conditions. And I have argued that Fischer-style cases also fail, for they are not cases where the agent could not have done otherwise. I shall conclude with a few remarks about the Principle.

Frankfurt noted at the beginning of his 1969 paper that the Principle 'has generally seemed so overwhelmingly plausible that some philosophers have even characterized it as an *a priori* truth' [Frankfurt 1969: 829]. Largely as a result of Frankfurt's influence, it would certainly be false to say that the Principle is now regarded as overwhelmingly plausible, let alone an *a priori* truth. But I want to suggest nonetheless that the Principle is indeed overwhelmingly plausible and explain why.

The Principle says that 'a person is morally responsible for what he has done *only if* he could have done otherwise' [Frankfurt 1969: 829, my italics]. So the Principle offers to capture a *necessary* condition for moral responsibility. The question is whether that is a plausible claim: whether it is plausible to hold that the possibility of doing otherwise is a necessary condition for moral responsibility.

The answer is that it is, because it seems unreasonable to hold someone morally responsible for something over whose occurrence he had no control: we do not hold people morally responsible for being born blind, blonde, or British, for instance. Nor do we hold people morally responsible for occurrences that they neither cause nor could have done anything to prevent.

Now, if a person does something and he could not have refrained from doing it, then it was not up to him whether he did that: he had no control over whether he did it or not.[21] And so it seems unreasonable to attribute

[21]I am, of course, claiming that moral responsibility requires what Fischer and Ravizza call 'regulative control' as opposed to merely 'guidance control' (see [Fischer and Ravizza 1998]). They themselves reject this view on the grounds that counterexamples to the Principle (either Frankfurt-style cases or Fischer-style cases) show that regulative control is not necessary for moral responsibility. But if my arguments are right, those cases do not show that.

any moral responsibility to him for doing it. And that is the truth that the Principle captures: when the 'could have done otherwise' condition is not met, questions about the extent of an agent's moral responsibility for what he did do not even arise.[22]

However, when the condition is met, it would be wrong to think that the extent of a person's moral responsibility is *settled* and *explained* by the mere fact that the person could have acted otherwise. The factors that settle and explain the extent of a person's moral responsibility over something he did are things such as what the agent wanted and intended to do, his motives in acting, how aware he was of what he was doing, whether there was provocation, influence, or coercion of any kind and if so how severe, etc.[23] And what the Principle tells us is that these issues arise only when the agent could have done otherwise. So, although the knowledge that a person could have acted otherwise does not tell us anything about why he did what he did, it does tell us that what he did was indeed something about which questions of the extent of his moral responsibility can be pertinently raised—questions which would not be pertinent, had that modal condition not been met. But, of course, when those questions are pertinent, the questions themselves go well beyond the issue whether the person could have avoided doing what he did.[24]

University of Southampton

Received: November 2006
Revised: July 2007

[22]The same is also true, *mutatis mutandis*, of omissions. So, suppose, for instance, that Jones decides 'for reasons of his own' not to save Smith. One might think that it is possible to construct a Frankfurt-style case where, had Jones been about to decide to save him, Black would have prevented this. If so, this appears to be a case where Jones is morally responsible for failing to save Smith but where he could not have 'acted otherwise', that is, where he could not have done anything other than fail to save him. So this appears to be a counterexample to the Principle for omissions.

But that appearance is deceptive and the case is not a counterexample either. For 'omission' is a normative concept, so that not everything that one doesn't do is an omission, but only those things one *ought* to do: my not singing an aria at the beginning of every lecture is not an omission (see [Alvarez 2001]). So, Jones's not saving Smith would be an omission only if Jones *ought* to have saved Smith. But if, as seems plausible (though this is also debated), 'ought' implies 'can', then we cannot say that Jones ought to have saved Smith for, given the background conditions, Jones could not have *saved* Smith (and the same goes for trying to save him, if we accept that, given the conditions, he could not have tried). So Jones is not morally responsible for failing to *save* Smith because, since that is not something he could have done, it is not the case that he ought to have done it. (So, to some extent, I agree with Frankfurt [1994], where he argues against the claim that there is an asymmetry between actions and omissions, as defended e.g. by J. Fischer and M. Ravizza [1991]. However, unlike Frankfurt, I think that, both for actions and omissions, moral responsibility *requires* the possibility of doing otherwise.)

This does not mean that Jones is wholly exonerated in these conditions, for he *is* blameworthy for *deciding* to not to save Smith. Again, one may think that this is enough to construct a counterexample, for one might argue that Jones is morally responsible for having *decided* not to save Smith even though he could not have 'done otherwise'. But the discussion of Fischer-style cases shows that this is not so. For, although it is true that Jones could not have *decided* to save Smith, he could have *refrained* from deciding *not* to save him. And, since it was up to him whether he decided not to save Smith or whether he refrained from taking that decision, it follows that he *could have done otherwise*.

[23]It is indeed issues like these, rather than the impossibility of doing otherwise, that explain why, for instance, to the extent that they are not aware of what they are doing, a somnambulist or a hypnotized person is not morally responsible for what he does.

[24]This paper was written during my tenure of a Mind Fellowship, and versions of it were presented at research seminars at several philosophy departments in the UK. I would like to thank the Mind Association, participants in those seminars, and especially Kadri Vihvelin, my colleague Aaron Ridley, and three referees for the *Australasian Journal of Philosophy* for their comments on earlier versions.

References

Alvarez, Maria 2001. Letting Happen, Omissions and Causation, *Grazer Philosophische Studien* 61: 63–81.

Aquinas, Thomas 1960–73. *Summa Theologiae*, ed. Thomas Gilby, Cambridge: Blackfriars.

Aristotle 1984. *The Complete Works of Aristotle*, ed. Jonathan Barnes, Princeton NJ: Princeton University Press.

Cain, James 2003. Frankfurt Style Cases, *Southwest Philosophy Review* 19: 221–9.

Fischer, John M. 1994. *The Metaphysics of Free Will*, Oxford: Blackwell.

Fischer, John M. 1999. Recent Work on Moral Responsibility, *Ethics* 110: 93–139.

Fischer, John M. 2003. Responsibility and Agent-Causation, in *Moral Responsibility and Alternative Possibilities*, ed. David Widerker and Michael McKenna, Aldershot, UK: Ashgate: 235–50.

Fischer, John M. and Ravizza, Mark 1991. Responsibility and Inevitability, *Ethics* 101: 258–78.

Fischer, John M. and Ravizza, Mark 1998. *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Frankfurt, Harry 1969. Alternate Possibilities and Moral Responsibility, *Journal of Philosophy* 66/4: 829–39.

Frankfurt, Harry 1978. The Problem of Action, *American Philosophical Quarterly* 15: 157–62.

Frankfurt, Harry 1982. What We Are Morally Responsible For, in *How Many Questions? Essays in Honor of Sidney Morgenbesser*, ed. L. S. Cauman et al. Indianapolis: Hackett.

Frankfurt, Harry 1994. An Alleged Asymmetry Between Actions and Omissions, *Ethics* 104: 620–3.

Geach, Peter 2000. Intention, Freedom and Predictability, in *Logic, Cause and Action: Essays in Honour of Elizabeth Anscombe*, ed. Roger Teichmann, Cambridge: Cambridge University Press: 73–81.

Ginet, Carl 1996. In Defence of the Principle of Alternate Possibilities: Why I don't Find Frankfurt's Argument Convincing, *Philosophical Perspectives* 10: 403–17.

Hart, H.L.A. 1968. *Punishment and Responsibility*, Oxford: Oxford University Press.

Hornsby, Jennifer. 1980. *Actions*, London: Routledge & Kegan Paul.

Hunt, D. P. 2000. Moral Responsibility and Unavoidable Action, *Philosophical Studies* 97: 195–227.

Kenny, A. J. P. 1975. *Will, Freedom and Power*, Oxford: Blackwell.

Kenny, A. J. P. 1989. *The Metaphysics of Mind*, Oxford: Oxford University Press.

Mele, Alfred R. 1992. *Springs of Action*, New York: Oxford University Press.

Mele, Alfred R. and David Robb 1998. Rescuing Frankfurt-style Cases, *The Philosophical Review* 107/11: 97–112.

Pereboom, Derk 2003. Source Incompatibilism and Alternative Possibilities, in *Moral Responsibility and Alternative Possibilities*, ed. David Widerker and Michael McKenna, Aldershot, UK: Ashgate: 185–200.

Reid, Thomas 1969 (1788). *Essays on the Active Powers of Man*, ed. Baruch A. Brody, Cambridge Mass.: MIT Press.

Vihvelin, Kadri 2000. Freedom, Foreknowledge, and the Principle of Alternate Possibilities, *Canadian Journal of Philosophy* 30: 1–24.

Widerker, David and Michael McKenna 2003. *Moral Responsibility and Alternative Possibilities*, Aldershot, UK: Ashgate.

Von Wright, G.H. 1963. *Norm and Action*, London: Routledge & Kegan Paul.