forthcoming in *Philosophical Explorations*

Frankfurt vs. Frankfurt: a new anti-causalist dawn

Ezio Di Nucci

In this paper I argue that there is an important anomaly to the causalist/compatibilist paradigm in the philosophy of action and free will.[1] This anomaly, which to my knowledge has gone unnoticed so far, can be found in the philosophy of Harry Frankfurt. Two of his most important contributions to the field – his influential counterexample to the Principle of Alternate Possibilities (henceforth *the Principle*) from *Alternate Possibilities and Moral Responsibility* (1969), and his 'guidance' view of action (1978) – are incompatible.

Frankfurt's counterexample to the Principle works only if we do NOT understand action as Frankfurt does in his guidance account. If, on the other hand, we understand agency in terms of the agent's guidance – as Frankfurt proposed in *The Problem of Action* (1978) – then his counterexample to the Principle fails because, then, counterfactual scenarios of Frankfurt-type counterexamples are such that what happens does not count as the relevant agent's action. So Frankfurt-type counterexamples do not show that the agent could not have avoided acting as she did: so they fail to offer a scenario in which the agent is intuitively responsible even though she could not have avoided acting as she did. Therefore Frankfurt-type counterexamples do not challenge the Principle, according to which "a person is morally responsible for what he has done only if he could have done otherwise" (1969, p. 829).

The importance of this inconsistency goes far beyond the issue of coherence within

[1] Thanks for useful comments to Neil Roughley and to three anonymous referees for this journal. And thanks to the audience of the Philosophisches Kolloquium, Universität Duisburg-Essen for a good discussion.

Frankfurt's philosophy. I shall argue that this inconsistency represents an important anomaly within the causalist/compatibilist framework; so that we should start to seriously consider having to move on from the established paradigm. First I am going to present Frankfurt`s two contributions and show that they are incompatible, then I will argue that this incompatibility poses a crucial challenge to the causalist/compatibilist paradigm in the philosophy of action and free will.

To be sure, by causalist/compatibilist paradigm I mean, on the one hand, the causal theory of action, which since Davidson's trailblazing *Actions, Reasons, and Causes* (1963) has established itself as the overwhelmingly dominant account of action explanation; and, on the other hand, compatibilism about free will and determinism. Most naturalists working in this field endorse both a version of the causal theory of action and a version of compatibilism. This is because, from a naturalistic point of view, those two go well together. Both offer ways in which rational autonomous human agency can be reconciled within nature: we can appeal to a kind of causal explanation in order to explain human agency (causal theory of action); and we need not renounce a causally deterministic view of the universe in order to acknowledge the autonomous character of human agency. Not only, then, both views fit naturalism; but they also fit each other well.

This cozy framework, I argue here, is challenged by Frankfurt`s inconsistency. In section 1 I present Frankfurt's counterexample to the Principle. In section 2 I present his guidance view of action. In section 3 I show that the two are incompatible. And in section 4 I argue that this incompatibility poses a challenge to the current paradigm.

1. *Counterexample to the Principle*

In *Alternate Possibilities and Moral Responsibility* (1969), Frankfurt offers a now famous counterexample to the Principle which goes 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 to him (Black is an excellent judge of such things) that Jones is going to decide to do something other than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. 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. In that case, it seems clear, Jones will bear precisely the same moral responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he do it (1969, pp. 835–36).

In this scenario, Jones freely and autonomously performs a certain action, in such a way that it appears obvious to us that Jones is responsible for performing that action. But Jones could not have done otherwise than performing the action in question. Therefore this appears to be a case in which Jones is morally responsible even though he could not have done otherwise – which contravenes the Principle.

The literature on this topic is enormous (see Fischer 1999 for a useful survey), but here I am primarily interested in Frankfurt´s original formulation of the counterexample. The idea is that either Jones, "for reasons of his own, decides to perform and does perform the very action Black wants him to perform" (1969, p. 836) or "Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do" (p. 835). Whichever the disjunct, Jones performs the action in question. This is crucial, because it is on this that the claim that Jones could not have done otherwise depends. In both the actual and the counterfactual scenario Jones does the action in question. Therefore when Jones does perform that action in the actual scenario, he could not have done otherwise than that action because he would have performed *that* action in the counterfactual scenario too.

Take the action of 'killing Smith': either Jones, 'for reasons of his own, decides to kill Smith and does kill Smith' or 'Black takes effective steps to ensure that Jones decides to kill Smith, and that he does kill Smith'. It follows that in both the actual scenario and the counterfactual scenario Jones kills Smith. Therefore when Jones kills Smith in the actual scenario, Jones could not have done otherwise than killing Smith. Therefore if we consider Jones responsible for killing Smith, then Jones is responsible for killing Smith despite the fact that Jones could not have done otherwise than killing Smith. So the Principle is false and moral responsibility does not require alternate possibilities.

One important question here is how to read the 'could have done otherwise' part of the Principle. We should not read that simply as requiring that it must have been possible for the

agent not to have done the action in question. Namely, the Principle does not just say that an agent is responsible for A-ing only if it was possible for the universe to unfold in such a way that she would not have A-ed. The counterfactual scenario in which the agent does not A must be suitably connected to the agent, so that whether or not the agent A-s is up to the agent. A simpler way of saying this is by saying that the Principle does not simply require the possibility that the agent does not A, but that it also requires that the agent would have been able to avoid A-ing. So the Principle would say that 'a person is morally responsible for what he has done only if he could have avoided doing that'.[2]

This clarification is important because otherwise the Principle is not interesting. Suppose Jones kills Smith. That the world could have unfolded in such a way that Jones would not have killed Smith – Planet Earth might have imploded the moment before Jones pulled the trigger (or the moment before Jones decided to kill Smith) – does not say anything about Jones's responsibility. It makes no difference to our intuitions or moral judgments. It is simply irrelevant. So if the Principle is to be in any way meaningful, then it must be read as 'a person is morally responsible for what he has done only if he could have avoided doing that'. And, importantly, this reading doesn't change the dialectic and intuitive appeal of Frankfurt's counterexample. Jones kills Smith freely and autonomously (for reasons of his own); so that we are inclined to say that Jones is responsible for killing Smith. But Jones could not have avoided killing Smith because Black would have taken effective steps to ensure that Jones decided to kill Smith and that he did kill Smith. So having specified how we ought to read the Principle does not affect Frankfurt's counterexample to it.

The other thing to emphasize about the dialectic of the counterexample is that it depends on the agent doing, in the counterfactual scenario, what she does in the actual scenario (e.g., killing Smith). It is because Jones kills Smith in both available scenarios that we say that Jones could not have avoided killing Smith, because he either kills Smith or he kills Smith. So the counterexample depends on Jones killing Smith in the counterfactual scenario. But this, again, ought not to be misunderstood. Imagine a version of Frankfurt's counterexample in which there are three, instead of two, possible scenarios:

1st scenario: Jones, for reasons of his own, decides to kill Smith and does kill Smith. Black does not intervene.

[2] On interpreting the Principle this way, see Alvarez 2009, Di Nucci 2010, and Di Nucci (forthcoming).

2nd scenario: Black takes effective steps to ensure that Jones decides to kill Smith and that he does kill Smith. As a result of this, Jones decides to kill Smith and does kill Smith.

3rd scenario: The moment before Jones decides to kill Smith (or the moment before Jones kills Smith), Planet Earth implodes. As a result of this, Jones does not decide to kill Smith nor does he kill him.

Suppose that the *1st scenario* ensues. Reasoning as before, we will say that Jones is responsible for killing Smith despite the fact that he could not have avoided killing Smith, and that therefore the Principle is falsified even in this case. But now the Principle has been falsified by a case in which it was possible for Jones not to kill Smith – as exemplified by the *3rd scenario*. Still, while in this case it is possible for the universe to unfold in such a way that Jones does not kill Smith, it is still impossible, it seems, for Jones to avoid killing Smith – since all that is opened to Jones (we are supposing) is either killing Smith autonomously, or killing Smith as a result of Black's intervention, or being vaporized by Planet Earth's implosion. But in none of these three scenarios does Jones avoid killing Smith. Therefore Jones is responsible for killing Smith even though he could not have avoided killing Smith. Therefore the Principle is false.

Returning to the original case, Jones kills Smith or he kills Smith. So in neither scenario does Jones avoid killing Smith. So Jones is responsible for killing Smith even though he could not have avoided killing Smith. Crucially, why could Jones not have avoided killing Smith? Because in the only other available scenario Jones does not avoid killing Smith – since he kills him. Not killing Smith does not imply avoiding killing Smith. But killing Smith does imply not avoiding killing Smith. Let us now turn to Frankfurt's discussion of guidance – I come back to the counterexample to the Principle in *Section 3*.

2. *Guidance*

In *The Problem of Action* (1978), Frankfurt argues against causal accounts of action (from Davidson 1963 onwards), proposing rather as a necessary and sufficient condition for agency the idea of *the agent's guidance*. Frankfurt challenges causalism – according to which some movement is an action only if it is caused by psychological states which rationalize it – mainly through cases of deviant causal chains, in which plainly accidental movements are caused by the very psychological states which rationalize them.

Frankfurt's account does not rely on the antecedents of actions, and it therefore does not depend on psychological states as the causes of action, as the causal theory does. On the other hand, it focuses on the relationship between an agent and her action at the time of action: "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. Moreover, the question of whether or not movements occur under a person's guidance is not a matter of their antecedents" (Frankfurt 1978, p. 45). Frankfurt initially distinguishes between two kinds of purposive movement (p. 46): purposive movements which are guided by the agent, and purposive movements which are guided by some mechanism that cannot be identified with the agent. Through the idea of purposive movement, Frankfurt gives us an insight into what the agent's guidance is:

> Behaviour is purposive when its course is subject to adjustments which compensate for the effects of forces which would otherwise interfere with the course of the behaviour, and when the occurrence of these adjustments is not explainable by what explains the state of affairs that elicits them. The behaviour is in that case under the guidance of an independent causal mechanism, whose readiness to bring about compensatory adjustments tends to ensure that the behaviour is accomplished. The activity of such a mechanism is normally not, of course, guided by us. Rather it is, when we are performing an action, our guidance of our behaviour (1978, pp. 47-48).

For some movement to be under the agent's guidance, then, the adjustments and compensatory interventions don't need to be actualized; it is just a question of the agent being able to make those adjustments and interventions: "whose readiness to bring about compensatory adjustments tends to ensure that the behaviour is accomplished" (ibid.). This latter point finds confirmation in Frankfurt's famous car scenario, where he stresses that guidance does not require those adjustments and interventions to take place; it only requires that the agent be able to make those:

> A driver whose automobile is coasting downhill in virtue of gravitational forces alone might be satisfied with its speed and direction, and so he might never intervene to adjust its movement in any way. This would not show that the movement of the automobile did not occur under his guidance. What counts is that he was prepared to intervene if necessary, and that he was in a position to do so more or less effectively. Similarly, the causal mechanisms which stand ready to affect the course of a bodily movement may never have occasion to do

> so; for no negative feedback of the sort that would trigger their compensatory activity might occur. The behaviour is purposive not because it results from causes of a certain kind, but because it would be affected by certain causes if the accomplishment of its course were to be jeopardized (1978, p. 48).

So guidance is a form of agential control over one's movements, such that only in the presence of this particular kind of control can agents be said to act. This form of control does not involve any *further* activity. Agents don't need to be doing anything in order to have guidance over their movements. But they do need to be in a position to intervene. So while actual intervention isn't necessary for guidance, the ability to intervene is. An agent's guidance can therefore be passive, as in the coasting scenario above. I will now show that understanding action in these terms, as Frankfurt proposes to do, *blocks* his counterexample to the Principle.

3. *Guidance and the counterfactual scenario*

Understanding action in terms of the agent's guidance, what happens in the counterfactual scenario of Frankfurt's counterexample to the Principle does not count as Jones's acting. But if Jones does not act in the counterfactual scenario – if, for example, what happens does not count as Jones killing Smith – then it is not true that when Jones kills Smith, either he kills Smith or he kills Smith – because in the counterfactual scenario Jones would not kill Smith, if we understand action in terms of guidance. In this section I am going to argue for the above, and I am going to show that if it is true that Jones does not kill Smith in the counterfactual scenario, then Frankfurt's counterexample fails.

Recall what Frankfurt says about the counterfactual scenario: "Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do" (1969, p. 835). Frankfurt characterizes these 'effective steps' as follows: "let Black manipulate the minute processes of Jones's brain and nervous system in some more 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" (1969, pp. 835-36). In Fischer's standard re-formulation, Black installs a mechanism in "Jones's brain which enables Black to monitor and control Jones's activities" (1982, p. 26).

Given the kind of direct control that Black is exercising upon Jones in the counterfactual

scenario, it doesn't look as though Jones can be said to be acting. Jones has no control over what his mind and body do, because Black is manipulating them to do exactly what *he* wants them to do. Jones is no more than an instrument in Black's hands – with no wishes of his own nor the ability to fulfill them. In the counterfactual scenario, Jones is no agent – if there is an agent, that can only be Black. This much is, it seems to me, intuitive. To treat Jones, rather than Black, as the agent in the counterfactual scenario, would amount to blaming the baton rather than the policeman. Further, it is not even as if Black and Jones are acting together: it is just Black who is *using* Jones, rather like the policeman uses the baton.

Now what's important here is not even how intuitive the above might sound: rather it is crucial that the claim that Jones is not acting in the counterfactual scenario follows from Frankfurt's idea of the agent's guidance.[3] Jones has no guidance over his behaviour in the counterfactual scenario: he does not control it; he cannot prevent it, regulate it, adjust it, or inhibit it. He cannot do any of the things that characterize the agent's guidance according to Frankfurt. And therefore he has no guidance over killing Smith. And therefore killing Smith is not his action because he hasn't got guidance over it. Indeed, killing Smith still looks likely to qualify as an action, but not Jones's action – rather Black's, if anybody.[4] Jones, according to Frankfurt's view of agency, does not act in the counterfactual scenario because what happens isn't under Jones's guidance. Therefore, in the counterfactual scenario, Jones does not kill Smith.

Others have already voiced this idea. Fischer: "But if Black's computer were to intervene, it is not clear that Jones would be acting (in the relevant sense) at all. Were Black's computer to intervene and directly manipulate Jones's brain state, we might say that Jones's bodily movements would not in the appropriate sense be his actions (or actions at all)" (1982, p. 31).

[3] The distinction between the intuitiveness of the claim and its following from Frankfurt's conception of agency is important in another respect: that the claim is intuitive, and that Frankfurt's conception of action captures it, is a consideration in favour of Frankfurt's view of agency. And, conversely, it is also a consideration against those causal views that Frankfurt criticizes. It is not at all clear that causal views could account for what happens in the counterfactual scenario not being Jones's actions. Neither control (or guidance) nor the origin of the agent's psychological states matter within causalism, so that if the rationalizing psychological states cause the relevant behaviour, then that behaviour counts as action. So the counterfactual scenario of Frankfurt-type cases might indeed constitute a counterexample to causalism (more on this in section 4). Finally, by accepting that it is simply intuitive that Jones does not act in the counterfactual scenario, the argument of this paper can be generalized: here I would not just be showing an inconsistency within Frankfurt's philosophy, but also offering a general argument against Frankfurt-type counterexamples to the Principle.

[4] Here I can't get into the interesting issue of what a guidance view would say about the actual scenario of Frankfurt-type counterexamples. Briefly, it seems that, according to the guidance view, Black participates to proceedings even in the actual scenario, where Jones and Black act before – because Black retains the capacity to directly intervene, correct, or inhibit Jones's performance.

McKenna: "when an agent's actions, or the deliberative machinery of her actions, are brought about by reasons independent of the agent's own rational machinery, then the actions, or the deliberative machinery are not hers. This is not to say that it is *a priori* impossible for an intervener to cause an agent to act, or even to cause her to have such and such deliberations. It is only to say that the actions or the deliberations are not hers" (1997, p. 83).

Here it might be objected that Black's intervention need not be characterized as strongly as I have: the way I have described the counterfactual scenario, Black takes over in such a way that Jones no longer has any control over what happens in the counterfactual scenario – Jones is rather like the pistol or the baton; and that, I have claimed, is incompatible with Frankfurt's guidance view of action. But what if Black's intervention is more limited? What if, for example, all Black does is twitching Jones's neural processes the moment before he decides whether to kill Smith or not; so that Black makes Jones decides to kill Smith, and then his job is done: from then on it is Jones who truly does the rest of the work – and one might imagine that the killing of Smith is more complicated than a mere trigger-pulling. So that it is clear that Jones has to exercise a lot of guidance in the process of killing Smith. And, this alternative reading goes, at these latter stages Black isn't doing anything – if not maybe monitoring that Jones does indeed kill Jones.

We can describe this alternative reading of the counterfactual scenario in terms of this useful metaphor: we don't need to suppose that Jones is driving, and that when Black realizes that Jones might go a different way from where Black wants to go, then Black takes over control of the wheel from Jones and drives the rest of the way to his desired destination. We might just suppose that when Black realizes that Jones isn't going to go towards the desired destination - say Jones is about to miss the crucial exit – Black grabs the wheel just long enough to direct the car towards the exit. And then leaves Jones to drive for the rest of the way once the car has been put on the right road by his intervention. The idea being that Black's intervention might have made Jones's driving less free, but Jones is still driving – he is still acting, because he is acting after Black's intervention.

I think we ought not to read the counterfactual scenario in the way proposed above, and that we should rather stick to my original reading. First, we must remember why Black has to intervene in the first place: it is because Jones is about to not kill Smith. So if Black's intervention is limited, and then it is again Jones who is 'at the wheel', then we ought to

suppose that Jones will use his regained control to not kill Smith – that is, after all, what he was going to do in the first place before Black's intervention. In order to prevent this reply, the supporter of the 'limited intervention' reading would have to stipulate that even though Black's intervention is limited, it is enough to guarantee that Jones will indeed kill Smith. But how can it guarantee that? Well, the most obvious way would appear to be my original reading: Black takes over.

Alternatively, one might suppose that Black operates his limited intervention to put Jones on the right road, and then keeps watch to makes sure that Jones does not do anything silly like reverting back towards his original destination – not killing Smith. But now the counterfactual scenario is just like the actual scenario: if Black does not have to intervene, we will think of Jones as responsible. So it is not just that Jones acts – he is intuitively responsible for what he does even though he could not have done otherwise because Black was keeping watch. This reading also violates another stipulation of Frankfurt's counterexample: that there are only two possibilities, the actual scenario and the counterfactual scenario; because on this reading there are counterfactual scenarios to the counterfactual scenario. As many counterfactual scenarios, in fact, as there are ways in which Black might have to intervene to bring Jones back onto the 'right' road. And, if as of above, Jones will naturally tend to revert back to his original destination, Black will be intervening all the time: which is equivalent to my original reading – Black takes over and therefore Jones does not act.

So much for the claim that Jones does not act in the counterfactual scenario. I now turn to arguing that, given that Jones does not act in the counterfactual scenario, Frankfurt's counterexample does not work. In order for Frankfurt's scenario to count as a counterexample to the Principle, it must offer a case in which Jones is responsible for killing Smith even though he could not have done otherwise – or, as we are reading it, even though he could not have avoided killing Smith. So Frankfurt has to show that Jones could not have avoided killing Smith. If Frankfurt's scenario does not show that Jones could not have avoided killing Smith, then his counterexample fails. Frankfurt shows that Jones could not have avoided killing Smith, supposedly, by constructing a case in which Jones kills Smith in the actual scenario, and Jones would have also killed Smith in the counterfactual scenario. But now we have shown that Jones does not actually kill Smith in the counterfactual scenario (Black, if anybody, does).

A simple methodological point: given that the available scenarios are only two, and given that what Frankfurt needs to show is that Jones could not have avoided killing Smith, it is enough that Frankfurt shows that in neither scenario does Jones avoid killing Smith. If Frankfurt can do that, then given the limited scenarios, it follows that Jones could not have avoided killing Smith. Also, given that – as we have already said – killing Smith, as Jones does in the actual scenario, is incompatible with avoiding killing Smith, then all that is left to show for Frankfurt is that Jones does not avoid killing Smith in the counterfactual scenario. This is because if Jones does not avoid killing Smith in the counterfactual scenario, and Jones kills Smith in the actual scenario (and therefore does not avoid killing Smith in the actual scenario), then Jones does not avoid killing Smith in both of the available scenarios. And therefore Jones could not have avoided killing Smith.

Frankfurt's aim is clear: showing that Jones does not avoid killing Smith in the counterfactual scenario. And how he achieves that is also clear: by claiming that Jones kills Smith in the counterfactual scenario, which would imply that Jones does not avoid killing Smith. But since we have now shown that Jones does not kill Smith in the counterfactual scenario, now Frankfurt's argument for the claim that Jones does not avoid killing Smith in the counterfactual scenario has been negated. But since not killing Smith does not imply avoiding killing Smith, then the fact that Frankfurt's claim that Jones kills Smith in the counterfactual scenario has been negated – together with its consequence that Jones does not kill Smith in the counterfactual scenario – do not, alone, imply that Jones avoids killing Smith in the counterfactual scenario. But, importantly, the burden of proof is not on me to show that Jones avoids killing Smith in the counterfactual scenario. The burden of proof is on Frankfurt (and defenders of his counterexample) to show that Jones does not avoid killing Smith in the counterfactual scenario.

Here we return upon traditional ground in the free will debate. To establish whether in the counterfactual scenario Jones avoids killing Smith, we must look at a crucial feature of Frankfurt-type counterexamples: what triggers the counterfactual intervener's intervention (a feature that has received much attention in the literature; see Fischer 1999). A traditional response to Frankfurt-type counterexamples has always been to point out that the counterfactual intervener's intervention depends, in turn, on what the agent does or appears to be about to do: "[Black] does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something other than what he wants him to do.

If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do" (Frankfurt 1969, p. 835).

Black's intervention is therefore conditional on whether Jones is going to decide to do something other than what Black wants him to do (side-stepping the epistemological level by assuming, as Frankfurt does, that Black is an excellent judge of such things).[5] So Black intervenes only if Jones is going to decide to do something else. So here's what the counterfactual scenario looks like: Jones is considering whether to kill Smith; Jones is 'going to decide to do something else'; Black finds this out; Black intervenes on Jones's brain processes, ensuring that Jones 'decides to do, and that he does do', what Black wants him to do; Smith is killed by Jones's hands. Now, as we have said, because of Black's direct control over Jones's thoughts and moves, what happens does not count as Jones's agency (it counts as Black's agency if anything); so Jones does not kill Smith, even though Smith is killed by Jones's hands.

Now consider the following two points in conjunction: (1) Jones, prior to Black's intervention, was going to do something else, he was not going to kill Smith; (2) Jones does not kill Smith, since what happens does not count as his actions because Jones's thoughts and moves were being manipulated by Black. These two considerations taken together suggest that in the counterfactual scenario Jones avoided killing Smith. Smith was, indeed, killed; but it isn't Jones who did it; and, more importantly, the very reason why Black had to intervene is exactly that Jones was not going to kill Smith. There is, in the counterfactual scenario, a basic consistency between Jones's original intentions (that he was not going to kill Smith suggests that he didn't intend to kill him or at least that he had ultimately deliberated not to kill him despite having given serious consideration to the possibility of doing so) and what happens in the end: namely that Jones does not kill Smith. And this, I am suggesting, counts in favour of the idea that Jones avoided killing Smith. What else could we possibly ask of him, after all? He did all he could, to be finally overpowered by Black. And even then, he didn't do it in the end: Black did. Fair enough, Jones couldn't save Smith and, probably, had no inclination to do so either. But he didn't mean to do it and he didn't in the end do it. Shouldn't that count as (or at least *towards*) avoiding or refraining from killing Smith?

[5] Here I don't mean to suggest that Black is a 'conditional intervener' as opposed to a 'counterfactual intervener' (see Vihvelin 2000 for the distinction). My argument works in both cases.

Let's put that to the test, using the definition of 'avoiding A-ing (or: refraining from A-ing)' introduced earlier: was it up to Jones whether or not he killed Smith? It certainly looks as though it was: Black intervened, and so Jones ended up not killing Smith. Had Black not intervened, Jones wouldn't have killed Smith either. The only difference between the two alternatives being not whether Jones killed Smith – because he doesn't in either – but whether Smith is at all killed, since in the latter alternative Smith does die.[6] While he would not have died at all, presumably, had Black not intervened.

Supporters of Frankfurt often insist that Black's intervention does not necessarily depend on something that Jones does (see, for example, Fischer 1999): it might very well depend on something that rather *happens* to Jones; something that Jones cannot, in turn, prevent from happening. And therefore if what is required is that Black's intervention depends on Jones's agency (or, more simply: on something that Jones does) – so that we can say that it was truly up to Jones whether or not Black intervened – then that requirement would not be met. Frankfurt himself anticipated this line of reasoning: "We can imagine that Jones has often confronted the alternatives - A and B - that he now confronts, and that his face has invariably twitched when he was about to decide to do A and never when he was about to decide to do B. Knowing this, and observing the twitch, Black would have a basis for prediction" (1969, p. 835).

So Black's intervention depends on whether or not Jones's face twitches. And whether or not his face twitches is not up to Jones: he can't help it twitching. So, in turn, whether or not Black intervenes is not up to Jones either. But what we are trying to establish is whether, when Black intervenes, Jones avoids killing Smith. And whether Jones avoids killing Smith when Black intervenes depends, in turn, on two points: Jones can't be said to have avoided killing Smith if he does indeed kill Smith – but that he doesn't do. And Jones can't be said to have avoided killing Smith if his not killing him was not up to Jones himself. Supporters of Frankfurt will argue that since Black's intervention was not up to Jones, then not killing Smith was not up to Jones either. But this does not follow: because that Jones does not kill Smith isn't the consequence of Black's intervention. Certainly, Black intervenes and Jones does not kill Smith. But it's not as if Jones does not kill Smith *because* Black intervenes. Had Black not intervened, Jones would have done something other than killing Smith. This counterfactual is supported by Frankfurt's statement that 'Jones is going to decide to do something else'. And it

[6] Just to be clear: the two alternatives here are not the traditional actual scenario and counterfactual scenario, but whether Black intervenes or not when Jones is 'going to decide to do something else'.

is only because Black judges that 'Jones is going to decide to do something else' that Black does indeed intervene. Had Jones not been 'going to decide to do something else', Black would not have intervened.

So, importantly, whether Black's intervention depends on something that Jones does or merely on something that happens to Jones doesn't actually matter to this argument. So Frankfurt's supporters cannot use the 'twitch' argument to claim that Jones does not avoid killing Smith in the counterfactual scenario. And if it is not true that Jones does not avoid killing Smith in the counterfactual scenario, then it is not true that Jones could not have avoided killing Smith. And if it is not true that Jones could not have avoided killing Smith, then Frankfurt's scenario doesn't work as a counterexample to the Principle because it does not offer a case in which an agent is obviously responsible for A-ing even though she could not have avoided A-ing. So, if we accept Frankfurt's 'guidance' conception of agency, then the Principle of Alternate Possibilities is safe.

4. *A new (anti-causalist) dawn*

The incompatibility between Frankfurt's counterexample and his guidance view poses a dilemma. On the one hand, we have possibly the most influential compatibilist argument; on the other, a serious alternative to the dominant causal theory of action. If the Principle is false, then moral responsibility does not require alternative possibilities, and then maybe moral responsibility and determinism can be compatible. Another classic compatibilist move is indeed the one Frankfurt rejects in *The Problem of Action* (1978) with his guidance understanding of agency: namely, causalism. If we can understand the relationship between reasons and actions in causal terms, then we can explain human actions without appealing to a model other than the causal one; we can, for example, explain human action without appealing to metaphysically problematic versions of control and freedom. So causalism in the explanation of action and compatibilism about free will and moral responsibility appear to go hand in hand. Therefore there is more than one reason for compatibilists to choose Frankfurt's counterexample over his conception of agency. Symmetrically, there is more than one reason for libertarians to choose his conception of agency over his counterexample: not just upholding the Principle, but also a view of agency that does not understand actions in terms of their causes.

It should not be surprising, then, that the counterexample to the Principle – a classic *locus* of

compatibilism – and the guidance view, an anti-causalist proposal, are incompatible. What should be surprising is that those two can both be found in Frankfurt's philosophy. And this not just because they pose a question of consistency within Frankfurt's philosophy, but because of the central role played by Frankfurt in the field in the last forty years. It is this that I regard as the anomaly: not really the incompatibility in itself, but that this arises within the same author, and a very influential one at that.

There are two alternatives: one the one hand, one could try to defend the current paradigm by resolving the apparent dilemma. Even accepting my argument, one option is to reject Frankfurt's anti-causalist view of action; or, at least, to strengthen the causal theory of action from some of the weaknesses emphasized by Frankfurt's proposal. The vast literature on deviant causal chains ought to be seen in this light[7]: an attempt to protect causalism from deviant counterexamples (against its sufficiency) so that the causal view can be presented as a full-blown sufficient account of agency. Otherwise one could reject Frankfurt's counterexample to the Principle (see Fischer's survey article from 1999 for a good many examples of such attempts).

These two moves remain solidly within the paradigm: causalists defend the causal theory of action from deviant counterexamples; libertarians defend the Principle from Frankfurt-style counterexamples. The alternative is to look beyond this established paradigm. And I think that there are good reasons to do so: the first, obvious reason is the inconsistency illustrated in this paper, and the crucial fact that this inconsistency can be found in the work of the same influential philosopher. Secondly, there is an important asymmetry between the two views that, I have argued here, are incompatible. It emerges from my discussion that the claim that Jones does not act in the counterfactual scenario doesn't just follow from Frankfurt's account of guidance, but it is also the most intuitive way of describing the case. The counterexample fails, then, not just because of Frankfurt's guidance view, but because of our intuitions about agency more in general. And at the same time those very intuitions give us an important reason to choose guidance over causalism.

It is worth noting that the intuition according to which Jones in the counterfactual scenario cannot possibly be said to act because his movements are under Black's direct control – because Jones is being operated and maneuvered by Black from a distance rather like a child

[7] Some influential works on deviant causal chains: Davidson 1973; Brand 1984; Thalberg 1984; Bishop 1989; Mele & Moser 1994.

would operate a radio controlled toy car or a builder a crane – is the same intuition according to which the climber of Davidson's (1973, p. 79) original deviant scenario cannot be said to be intentionally letting go of his fellow climber because he merely loses his grip on the rope. What happens is an accident, not an action. The absence of control or guidance is, then, at the root of deviant counterexamples.

What Frankfurt's anomaly points to is the possibility of rejecting the causal theory of action while at the same time remaining firmly in naturalist, if not compatibilist, territory. That is also the sense in which Frankfurt's anomaly points to a paradigmatic shift: the idea is that one can reject the causal theory of action, and at the same time reject Frankfurt-style compatibilist arguments, without crossing over to libertarianism. This is indeed the direction in which Frankfurt's text points: "Despite its popularity, I believe that the causal approach is inherently implausible and that it cannot provide a satisfactory analysis of the nature of action. *I do not mean to suggest that actions have no causes*; they are as likely to have causes, I suppose, as other events are. My claim is rather that it is not part of the nature of an action to have a prior causal history of any particular kind" (1978: 42 – emphasis mine).

Frankfurt's reservations, then, aren't with whether actions have causes; but with whether the fact that something is an action can be explained by appeal to its causal history. A common worry is whether a certain causal history is sufficient for agency: and deviant causal chains are cases in which, supposedly, the right causal history (one comprising of rationalizing mental states) is not sufficient. But there have also been challenges to the necessity of rationalizing mental states: Dreyfus's skilled activity (1984, 1988, 2005); arational actions (Hursthouse 1991), emotional actions (Goldie 2000), passive actions (Zhu 2004), habitual actions (Pollard 2003 & 2006), omissions (Sartorio 2005 & 2009), and automatic actions (Di Nucci 2008).

The idea is that, often, when we act intentionally, we cannot point to mental states with the relevant rationalizing content. Here are some examples given by Hursthouse: rumpling someone's hair, "throwing an 'uncooperative' tin opener on the ground" (ibid, p. 58), jumping up and down in excitement, "covering one's face in the dark [out of shame]" (ibid), "covering one's eyes [in horror] when they are already shut" (ibid). There are no reasons why we do such things, not in the sense of goals anyhow. Nor are there intentions involved. Still, we do these things intentionally: those things are certainly actions of ours. The same problem arises

with skilled, habitual, and routine activities: think of driving, playing football, or making coffee in the morning. There might be a goal, but we don't actually have 'a goal in mind'. We often do those things without thinking, sometimes even without paying attention to our performance. That is, indeed, the point of skills and habits: we have perfected our performance and now we no longer need to waste cognitive resources on it.

There is another argument against causalism which I have anticipated in Section 3: if, as I have argued in this paper, Jones does not act in the counterfactual scenario, then causal theories of action have a problem; because they don't have the conceptual means to show that Jones does not act in the counterfactual scenario. According to causalism, some movement counts as an action if and only if it is caused in the right way by psychological states which rationalize it.[8] Given that Black manipulated Jones's psychological states so that Jones would intend to kill Smith, then according to causalism Jones's movements in the counterfactual scenario should be actions: they are caused, in the right non-deviant way, by psychological states which rationalize them – say an intention to kill Smith. But, we have argued, Jones does not kill Smith – Black does (I won't repeat here the argument from Section 3). Causalism fails to account for this because it does not have any control or guidance requirements, nor does it have any historical requirements. Namely, Black's peculiar place in the history of Jones's psychological states does not matter for causalism. Therefore causalism fails to explain why Jones's movements in the counterfactual scenario aren't actions.

Requiring for intentional agency that some movement be caused by rationalizing mental states, the traditional causal theory of action runs against the difficulties illustrated so far. These problems, together with deviant causal chains and with the control intuitions already discussed, suggest that we look beyond the causal theory of action. And, indeed, some version of Frankfurt's own account of guidance looks promising in accommodating both our control intuitions about action and the above problems with both the necessity and sufficiency of rationalizing mental states in the causal history of actions. Importantly, though, these problems appear to be independent from determinism and compatibilism. One can genuinely raise these objections to the causal theory of action without suggesting that actions are anything other than another natural component subject to the laws. Libertarian intuitions and libertarian arguments appear to be silent on the above concerns with the causal theory of action. Therefore it seems no coincidence that Frankfurt would raise some of those problems

[8] This is the formula used to side-step the deviant causal chains problem already emphasized in this section.

while at the same time coming up with a groundbreaking argument for compatibilism. It is not an issue of being charitable to Frankfurt's philosophy: rather, we are now starting to see how this inconsistency in his work *makes sense*. Naturalism does not impose us the causal theory of action, nor does determinism.

This paper is not the place to develop a full-blown alternative to causalism. So here I just want to touch upon two crucial points: firstly, if some concept of guidance, as I believe, should be part of this alternative account of agency – either by replacing or supplementing a causalist story; and if this alternative view is to be fully naturalistic, then the concept of guidance must not be understood in libertarian terms. Here there are two promising alternatives: one possibility is to develop such a view by going in the direction of Fischer and Ravizza's (1998, p. 31) *guidance control*. Alternatively, the capacity for intervention, correction, and inhibition that characterizes guidance according to Frankfurt's original formulation could be accounted for in terms of what has been recently labeled (by Clarke 2009) *New Dispositionalism*: in brief, the idea (put forward in different versions by Smith 2003, Vihvelin 2004, and Fara 2008) is that having a certain ability to act consists of or depends on having certain dispositions (depending on which of the above versions one takes). Unmanifested dispositions (finkish or masked dispositions) are compatible with determinism; therefore unexercised abilities are also similarly compatible.

Secondly, if guidance is to be developed into a full account of agency, it must be argued that guidance can be sufficient for agency, and not just necessary. If, then, guidance is to be a sufficient condition for agency, and guidance is to be independent from rationalizing mental states (otherwise we would again run into the problems already emphasized), then we would be offering an account of agency that does not directly appeal to the agent's motivation. Three things here: first, this conclusion might be too quick in overlooking externalism. Explaining agency without appealing to rationalizing mental states does not mean, according to externalists, explaining agency without appealing to reasons or motivation because, crudely put, reasons are facts rather than psychological states (see Stout 1996, Collins 1997, Dancy 2000, Alvarez 2010).

Second, this conclusion would similarly overlook what used to be called the *Logical Connection Argument* (Anscombe 1957, Hampshire 1959 Melden 1961, von Wright 1971) against which Davidson's (1963) original statement of the causal view was addressed. If the

relation between an action and the reason why that action is performed is rational, then it cannot be causal – that was the thrust of the old argument. Therefore denying that rationalizing mental states as causes are necessary for agency does not amount to denying the role of motivation simply because the motivational aspect does not entail the causal aspect; just as, in my previous point, the motivational aspect does not entail the psychological aspect.

But there is a third, wider point: is motivation actually necessary for agency? We can easily imagine scenarios in which agents have complete control over a certain movement but no motivation to perform that movement. In such cases, a minority would question the intentional nature of the act (that's the so-called Simple View, refuted by Bratman: see Bratman 1984 & Di Nucci 2009); others would question what implications the lack of motivation has on responsibility or permissibility (the so-called Principle of Double Effect). But as to whether that movement counts as an act, on that point control or guidance suffice.

Let us take stock: in this section I have argued that Frankfurt's incompatibility is an anomaly to the current causalist/compatibilist framework. I have explained why we should take this anomaly seriously, and shown in which direction it points to: a naturalistic view of agency centered on guidance.

Universität Duisburg-Essen
45117 Essen, Germany
ezio.dinucci@uni-due.de

References

Alvarez, M. (2009), 'Actions, Thought-Experiments, and the Principle of Alternate Possibilities', *Australasian Journal of Philosophy* 87/1: 61-81.

Alvarez, M. (2010), *Kinds of Reasons*. Oxford UP.

Anscombe, G.E.M. (1957), *Intention*. Basil Blackwell.

Bishop, J. (1989), *Natural Agency. An Essay on The Causal Theory of Action*. Cambridge University Press.

Brand, M. (1984), *Intending and Acting*. MIT Press.

Bratman, M. (1984), 'Two Faces of Intention', *Philosophical Review* 93: 375-405.

Clarke, R. (2009), 'Dispositions, Abilities to Act, and Free Will: The New Dispositionalism', *Mind* 118: 323-351.

Collins, A. W. (1997), 'The psychological reality of reasons', *Ratio*, X: 108-123.

Dancy, J. (2000), *Practical Reality*. Oxford UP.

Davidson, D. (1963), 'Actions, Reasons, and Causes', *Journal of Philosophy* 60: 685-700.

Davidson, D. (1973), 'Freedom to Act', in Honderich, T. (ed.), *Essays on Freedom and Action*. Routledge and Kegan Paul, 137-56.

Di Nucci, E. (2008), *Mind Out of Action*. VDM Verlag.

Di Nucci, E. (2009), 'Simply, false', *Analysis* 69/1: 69-78.

Di Nucci, E. (2010), 'Refuting a Frankfurtian Objection to Frankfurt-type Counterexamples', *Ethical Theory and Moral Practice* 13/2: 207-213.

Di Nucci, E. (forthcoming), 'Frankfurt counterexample defended', *Analysis*.

Dreyfus, H. (1988), 'The Socratic and Platonic Bases of Cognitivism', *AI & Society* 2: 99-112.

Dreyfus, H. (2005), 'Overcoming the Myth of the Mental: How Philosophers Can Profit from the Phenomenology of Everyday Expertise', *APA Pacific Division Presidential Address*.

Dreyfus, H. & Dreyfus, S. (1984), 'Skilled Behavior: The Limits of Intentional Analysis', in Lester, E. (ed.), *Phenomenological Essays in Memory of Aron Gurwitsch*. The University Press of America.

Fara, M. (2008), 'Masked Abilities and Compatibilism'. *Mind*, 117, pp. 843–65.

Fischer, J.M. (1982), 'Responsibility and Control', *The Journal of Philosophy* 79/1: 24-40.

Fischer, J.M. (1999), 'Recent Work on Moral Responsibility', *Ethics* 110/1: 93-139.

Fischer, J.M. & Ravizza, M. (1998), *Responsibility and Control*. Cambridge UP.

Frankfurt, H. (1969), 'Alternate Possibilities and Moral Responsibility', *Journal of Philosophy*

66: 829-39.

Frankfurt, H. (1978), 'The Problem of Action', *American Philosophical Quarterly* 15: 157-162.

Goldie, P. (2000), 'Explaining expressions of emotions', *Mind* 109: 25-38.

Hampshire, S. (1959), Thought and Action. Chatto and Windus.

Hursthouse, R. (1991), 'Arational Actions', *Journal of Philosophy* 88 (2): 57-68.

McKenna, M. (1997), 'Alternative Possibilities and the Failure of the Counter-Example Strategy', *Journal of Social Philosophy* 28 (3): 71-85.

Melden, A. I. (1961), *Free Action*. Routledge & Kegan Paul.

Mele, A. and Moser, P. K. (1994), 'Intentional Action', Nous 28: 39-68. Pollard, B. (2003), 'Can Virtuous Actions Be Both Habitual and Rational?', *Ethical Theory and Moral Practice* 6: 411-425.

Pollard, B. (2006), 'Explaining Actions with Habits', *American Philosophical Quarterly* 43: 57-68

Sartorio, C. (2005), 'A new asymmetry between actions and omissions', *Nous* 39: 460-482.

Sartorio, C. (2009), 'Omissions and Causalism', *Nous* 43: 513-530.

Smith, M. (2003), 'Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion', in Stroud, S. & Tappolet, C. (eds.) *Weakness of Will and Practical Irrationality.* Oxford UP.

Stout, R. (1996), *Things that happen because they should*. Oxford UP.

Thalberg, I. (1984), 'Do our intentions cause our intentional actions?', *American Philosophical Quarterly* 21: 249-260.

Vihvelin, K. (2000), 'Freedom, Foreknowledge, and the Principle of Alternate Possibilities', *Canadian Journal of Philosophy* 30: 1-24.

Vihvelin, Kadri 2004: 'Free Will Demystified: A Dispositional Account'. *Philosophical Topics*, 32, pp. 427–50.

von Wright, G.H. (1971), *Explanation and Understanding*. Cornell UP.

Zhu, J. (2004), 'Passive Action and Causalism', *Philosophical Studies* 119: 295-314.