# AGENT CAUSATION AND RESPONSIBILITY: A REPLY TO FLINT

Michael Bergmann

In an earlier paper I argued that if we help ourselves to Molinism, we can give a counterexample – one avoiding the usual difficulties – to the Principle of Alternate Possibilities:

PAP. A person is morally responsible for performing a given act only if she could have acted otherwise.

Thomas Flint has proposed three objections to my counterexample. In this paper I respond to each.

In my "Molinist Frankfurt-Style Counterexamples and the Free Will Defense,"[1] I argued that if we help ourselves to Molinism, we can give a counterexample – one avoiding the usual difficulties – to the Principle of Alternate Possibilities:

PAP. A person is morally responsible for performing a given act only if she could have acted otherwise.

In his "On Behalf of the PAP-ists: A Reply to Bergmann,"[2] Thomas Flint proposes three objections to my counterexample. In what follows, I respond to each.

## *I. Two Preliminary Matters*

Before considering Flint's three objections, it will be helpful if I begin by (i) explaining the way in which I use the term 'agent cause' (since it features so prominently in my counterexample and in my discussion of Flint's three objections) and (ii) briefly describing the counterexample I proposed.

I use the term 'agent cause' as follows:

AC. X is the agent cause of e iff each of the following three conditions is satisfied:
1. X is a substance that had the power to bring about e
2. X exerted its power to bring about e
3. nothing distinct from X (not even X's character) caused X to exert its power to bring about e.

This is a slight modification of the typical Reidian understanding of agent causation, which, according to William Rowe, is just like the above except that it replaces 3 with:

> 3*. X had the power to refrain from bringing about e.[3]

Rowe argues that 3* entails 3.[4] That seems plausible. And if it's true, then agent causation, as I understand it, seems to be weaker than the Reidian notion of agent causation insofar as my sort of agent causation seems to be entailed by the Reidian sort without entailing it.

With this understanding of agent causation in mind, let's consider my proposed counterexample. There is an agent, Jones, of whom the following subjunctive conditional of agent causation is true:

> A. If from t* up until t Jones were in circumstances K and Demon didn't take away Jones's powers at t with respect to V1 [a volition to pull the trigger of the gun in Jones's hand], then Jones would agent-cause V1 at t.

There is also a powerful being, Demon, with knowledge of subjunctive conditionals of agent causation such as A (here's where the Molinist component comes in). In addition to having such knowledge, Demon also knows that Jones will be in circumstances K from t* up until t. Since Demon knows all this and wants Jones to agent-cause V1 at t (i.e., to pull the trigger, which will result in Smith's death), Demon is happy *not* to take away Jones's powers at t with respect to V1. However, the following counterfactual is true of Demon:

> C. If A were false, Demon would know it (long before t) and would take away Jones's powers at t with respect to V1.

In fact, if A were false, not only would Demon take away Jones' powers at t with respect to V1, Demon would himself cause Jones to cause V1, thereby ensuring Smith's death. Finally, the antecedent of A is true (which, given the truth of A, implies that the consequent of A is true as well). I argued that in this example, Jones couldn't do otherwise than agent-cause V1 even though Jones is responsible for V1.[5]

## *II. Flint's First Objection*

Flint's first objection focuses on my claim that Jones is responsible for V1. To see exactly what his objection is targeting, it will be helpful to summarize the two points I made in defense of that claim. First, I noted that Jones would clearly be responsible for V1 if he agent-caused V1 when Demon wasn't present. This is because, if Jones agent-causes V1, the causal buck for V1 stops with Jones. Second, I claimed that changing the case to include Demon (in the way described in my counterexample) won't change anything relevant since, in the Demon case, Jones does exactly what he does in the Demon-less case and "he does so with absolutely no interference or influence from Demon".[6] It is this *second* point that Flint's objection targets. For, says Flint, the fact that Jones's powers to do otherwise than agent-cause V1 are removed when we add Demon to the situation "makes nonsense of the claim that there has been *no* interference or influence".[7] Perhaps I didn't state that second point as carefully as I should have. I didn't intend to say that there was no way in which the addition of Demon to the situation affects Jones. Clearly it does. My point was just that the addition of Demon to the situation results in no interference with or influence on *Jones's agent-causing of V1*.[8] In both the Demon case and the Demon-less case, the causal buck for V1 stops with Jones. That is sufficient for saying that Jones is responsible for V1.

### *III. Flint's Second Objection*

Flint's second objection is that my counterexample forces us to credit Demon with causing Jones to agent-cause V1 and that this is incoherent:

> Demon has set up a situation in which there's only one thing Jones can cause, and *where he can't refrain from exercising his power to cause it*. This surely seems like a situation in which Demon has caused Jones to exercise his power to agent-cause V1. And this even Bergmann would allow cannot be; his third condition of agent causation ... implies that nothing distinct from Jones could cause Jones to agent-cause anything. So Bergmann's counterexample may be incoherent even on his own account of agent causation.[9]

This objection relies quite heavily on the point that, in my counterexample, Jones "can't refrain from exercising his power to cause [V1]". But does my counterexample really entail that Jones can't refrain from exercising his power to cause V1? In the first half of his paper, Flint develops an argument for the conclusion that my counterexample does, in fact, have that consequence. Let's take some time to consider whether that argument succeeds.

According to Flint, there is a power I don't discuss in giving my counterexample, namely, the power to *nonintentionally* refrain from agent-causing V1, where this is distinct from the power to *intentionally* refrain from agent-causing V1.[10] (To intentionally refrain from agent-causing V1 involves agent-causing the relevant intention, something not involved in nonintentionally refraining from agent-causing V1.) Flint thinks that my counterexample entails that Jones lacks this power to nonintentionally refrain from agent-causing V1. For suppose that Jones had this power and exercised it. That is, suppose that:

> R. Jones refrains from agent-causing V1 at t.

Flint argues that R→~C[11] (where C is the counterfactual, mentioned earlier, that is true of Demon in my counterexample). But that shows that if Jones has the power to nonintentionally refrain from agent-causing V1, then "Jones has a power the exercise of which counterfactually implies the falsity of C".[12] And that conflicts with something I say about my counterexample, namely, that C would be true of Demon no matter which of his powers Jones exercised.[13] From this Flint concludes that my counterexample entails that Demon *takes away* Jones's power to nonintentionally refrain from agent-causing V1. That's why Flint says that, in my counterexample, Jones can't refrain from exercising his power to cause V1.

The reason I didn't mention this power to nonintentionally refrain from agent-causing V1 is that it isn't an active power that Jones can *exercise*.[14] Instead, it is the potential for something to happen to Jones. Some might call it a passive power – something that can be activated (by something distinct from Jones); but it isn't something that can be exercised by Jones. In fact, I don't even want to say it is a passive power to nonintentionally *refrain* from agent-causing V1. The term 'refrain' has connotations of being intentional. I prefer to think of it as Jones's potential to nonintentionally *fail* to agent-cause V1.

Now suppose that I grant (as I do) that Jones has this potential to nonintentionally fail to agent-cause V1. Can Flint's argument (that R →~C) be used to show that I'm in trouble? No. For now our question isn't "What would happen if Jones were to exercise his power to nonintentionally refrain from agent-causing V1?" Instead, the question is "What would happen if Jones's potential to nonintentionally fail to agent-cause V1 were activated?" In other words, the question is "What would happen if

> R*: Jones nonintentionally fails to agent-cause V1 at t

were true?" Thus, in order to show that I'm in trouble when I allow that Jones has the potential to nonintentionally fail to agent-cause V1, Flint must argue that R* →~C.

But once we replace R in Flint's argument with R*, we can see where it goes wrong. As Flint makes clear when he lays out the argument, it depends for its success on the validity of the following line of reasoning:

> Jones could exercise his power to refrain only if Demon has not taken away that power. And if Demon didn't take away that power, then obviously he didn't take away all of Jones's powers. So
>
> (ii) R → G1 [where G1 is the claim that Demon didn't take away *all of* Jones's powers at t with respect to V1].[15]

Thus, unless R → G1, Flint's argument fails to show that R → ~C. And since our focus is R* rather than R, we may conclude that, unless R* → G1, Flint's argument can't be used to show that R* → ~C. And the fact is that R* does *not* counterfactually imply G1.

To see why, consider once again the following counterfactual which is true of Demon in my counterexample:

> C. If A were false, Demon would know it (long before t) and would take away Jones's powers at t with respect to V1.

When I said (in stating the consequent of C) that Demon would take away (all of) Jones's powers at t with respect to V1, I was speaking of all of the *active* powers with respect to V1 that Jones can *exercise*. I was not speaking of the potentialities Jones has with respect to V1 – the things that can happen to him with respect to V1. This was made evident in my description of the example when I said that Demon would take away all of Jones's powers at t with respect to V1 *and that Demon himself would cause Jones to cause V1 at t*.[16] For since, in that counterfactual scenario, Demon caused Jones to cause V1 at t, it is obvious that Jones still had the potential to be caused to cause V1 despite the fact that all his powers with respect to V1 had been taken away. So clearly I wasn't suggesting that Demon was taking away all of Jones's *potentialities* with respect to V1; it was *active* powers with respect to V1 that I had in mind. Once we see this, it is easy to see that R* doesn't counterfactually imply G1 (where G1 is understood to be speaking of active powers). For the fact that Jones has the *potential* to nonintentionally fail to agent-cause V1 doesn't counterfactually imply that Jones is left with some *active* powers with respect to V1 that he can exercise. Once we see that R* doesn't counterfactually imply G1, we can (as I noted earlier) conclude that Flint's argument can't be used to show that R* → ~C. And from that we may conclude that it can't be used to show that my counterexample is inconsistent with the claim that Jones has the potential to nonintentionally fail to agent-cause V1.

Now all of this is, of course, directly relevant to Flint's second objection. For as we noted earlier, this objection relied on Flint's claim that, in my counterexample, Jones can't *refrain* from exercising his power to cause V1. If refraining from exercising the power to cause V1 can include such things as nonintentionally failing to agent-cause V1, then Flint has failed to establish that claim. If, on the other hand, it *can't* include such things, then we can note that even if Jones can't *refrain* from exercising his power to cause V1, he can nonintentionally *fail* to agent-cause V1. Either way the objection fails. For Flint wanted to suggest that, in my counterexample, Jones was *forced* to agent-cause V1. He created this impression by arguing that, according to my counterexample, Jones couldn't even nonintentionally fail to agent-cause V1. But as we have seen, his argument does not establish that conclusion.[17]

## *IV. Flint's Third Objection*

Let's look briefly at the third and last of Flint's three objections. According to this objection, my counterexample is incoherent if one understands agent causation in the way Thomas Reid and most agency theorists understand it (where clause 3 from my definition of agent causation is replaced with 3*), rather than in the way I understand it.[18] The problem, says Flint, is that my example entails that Jones agent-causes V1 even though he lacks the power to refrain from causing V1. And yet, if we insist upon clause 3* instead of merely clause 3, then agent-causing V1 requires the power to refrain from causing V1. So, says Flint, on the Reidian account of agent-causation, my example is incoherent.

I have two responses. First, it is difficult to see why this observation (even if correct) would cause any problem for me. When I described my counterexample, I made it clear that I was thinking of agent causation as entailing the satisfaction of clause 3, not clause 3*. Why is it a problem for me that my counterexample *would* be incoherent if understood in a way that I explicitly said it was *not* to be understood? Perhaps the suggestion is that I shouldn't be using the term 'agent causation' in a non-Reidian way. I disagree. I think that the account of agent causation I gave does a better job than Reid's account of capturing what at least *some* philosophers have in mind when thinking about agent causation.

Second, I'm not sure that my counterexample would be incoherent even if agent causation were understood in the Reidian way. It all depends on how we are to understand:

3*. X had the power to refrain from bringing about e.

Must X's power to refrain be thought of as an active power that X can exercise? If so, then my example would be incoherent given a Reidian account of agent causation. Or can the power to refrain from bringing about e be merely a passive power to have something happen to one? Can it, for example, be merely the potential to nonintentionally fail to agent-cause e? If so, then, in light of the remarks I made in response to Flint's second objection, I don't see why my example would be incoherent even given a Reidian account of agent causation.[19]

*Purdue University*

NOTES

[1] *Faith and Philosophy* 19 (2002), 462-78.

[2] Ibid., pp. 479-84.

[3] See William Rowe, "The Metaphysics of Freedom: Reid's Theory of Agent Causation," *American Catholic Philosophical Quarterly* LXXIV (2000), 427.

[4] This is what his response to objection II on p. 430 of "The Metaphysics of Freedom" amounts to.

[5] See section 2.2 of "Molinist Frankfurt-Style Counterexamples" where I lay out this example in a little more detail.

[6] "Molinist Frankfurt-Style Counterexamples,"p. 468.

[7] "On Behalf of the PAP-ists," p. 482.

[8] That's what I meant when I said *"he does so* with absolutely no interference or influence from Demon".

[9] "On Behalf of the PAP-ists," p. 482 (emphasis is mine).

[10] "On Behalf of the PAP-ists," p. 481.

[11] I follow Flint in using the single-line arrow (→) to represent counterfactual implication.

[12] "On Behalf of the PAP-ists," p. 482.

[13] See premise 16 on p. 471 of my "Molinist Frankfurt-Style Counterexamples".

[14] So I don't think it makes sense for Flint to ask, as he does, of this power "What would have happened if Jones were to exercise it?" (See his "On Behalf of the PAP-ists," p. 481.)

[15] "On Behalf of the PAP-ists," p. 481.

[16] In my original paper I said of Demon that "if he knows that Jones won't

agent-cause V1 at t, he will intervene as follows: he will take away Jones's powers at t with respect to V1 and will cause V1 himself at t." (See "Molinist Frankfurt-Style Counterexamples," p. 467.)

17. It's worth noting here that weakly actualizing a state of affairs needn't involve causing it to obtain. Thus, even if one could show that the state of affairs *Jones's agent-causing V1* were weakly actualized by Demon, that wouldn't be enough to show that Jones was forced to (or caused to) agent-cause V1. (To weakly actualize a state of affairs, S1, is to cause the actualization of a distinct state of affairs, S2, which is such that if it were actual, S1 would be actual. For an explanation of why weakly actualizing a state of affairs needn't involve causing it to obtain, see Plantinga's *Nature of Necessity* (Oxford: Clarendon Press, 1974), pp. 172-73.)

18. "On Behalf of the PAP-ists," pp. 482-83.

19. Thanks to Jeffrey Brower and William Rowe for comments on earlier drafts.