A version of this paper appears in *Philosophical Studies* **157** (2012), pp. 383-398.

AGENT CAUSATION AS THE SOLUTION TO ALL THE COMPATIBILIST'S PROBLEMS

Ned Markosian

Abstract: In a recent paper I argued that agent causation theorists should be compatibilists. In this paper, I argue that compatibilists should be agent causation theorists. I consider six of the main problems facing compatibilism: (i) the powerful intuition that one can't be responsible for actions that were somehow determined before one was born; (ii) Peter van Inwagen's modal argument, involving the inference rule (β); (iii) the objection to compatibilism that is based on claiming that the ability to do otherwise is a necessary condition for freedom; (iv) "manipulation arguments," involving cases in which an agent is manipulated by some powerful being into doing something that he or she would not normally do, but in such a way that the compatibilist's favorite conditions for a free action are satisfied; (v) the problem of constitutive luck; and (vi) the claim that it is not fair to blame someone for an action if that person was determined by forces outside of his or her control to perform that action. And in the case of each of these problems, I argue that the compatibilist has a much more plausible response to that problem if she endorses the theory of agent causation than she does otherwise. Keywords: agent causation, compatibilism, freedom and determinism.

1 Introduction

In an earlier paper I proposed an unusual variation on the traditional theory of agent causation.[1] On the view that I proposed, all that is required for *moral*

[1] Markosian, "A Compatibilist Version of the Theory of Agent Causation." For more on agent causation see for example Suarez, *Disputationes Metaphysicae*; Chisholm, "Human Freedom and the Self;" Taylor, *Action and Purpose*; van Inwagen, *An Essay on Free Will*; O'Connor, "Agent Causation" and *Persons and Causes*; Clarke, "Toward a Credible Agent-Causation Account of Free Will" and *Libertarian Accounts of Free Will*; and Turner and Nahmias, "Are the Folk Agent-Causationists?"

freedom (that is, the kind of freedom that is necessary for moral responsibility) is that the act in question be caused by its agent. In particular, it is not required that the act be in some way indeterministic. (Nor is it required that there be any indeterminism in the causal history of the act.) Here's the view.

> **The Compatibilist Version of the Theory of Agent Causation (COMTAC):** A is *morally free* iff A is caused by A's agent.

The absence of any other condition for an action's being morally free makes COMTAC a wholly compatibilist theory, since it means that an action can be free, according to COMTAC, even if the strongest form of determinism is true. This feature of COMTAC makes it unusual among theories of freedom that involve agent causation. For every previous version of the theory of agent causation has been an incompatibilist theory, requiring, for an action to be morally free, both (i) that it be caused by its agent and (ii) that there be some indeterminism in the causal history of the action.[2]

The main thesis of my previous paper was that the most plausible version of the theory of agent causation is a wholly compatibilist version of that theory, namely, COMTAC. So, in a sense, that paper was aimed at agent causation theorists, and was designed to convince them to become compatibilists. The present paper, on the other hand, is aimed at compatibilists, and is designed to convince them to become agent causation theorists. That is, the principal thesis of this paper is that compatibilists should be agent causation theorists.

My argument for this thesis is, essentially, that the main problems facing compatibilism can all be solved fairly easily by appealing to the phenomenon of agent causation. In order to show this, I consider six of the main problems facing compatibilism, and in the case of each of these problems, I argue that the compatibilist has a much more plausible response to that problem if she endorses the theory of agent causation than she does otherwise.

Before turning to the six problems for compatibilism, however, let me briefly explain the main motivation for COMTAC. Consider the following example.

---

[2] See for example Suarez, *Disputationes Metaphysicae*; Chisholm, "Human Freedom and the Self;" Taylor, *Action and Purpose*; van Inwagen, *An Essay on Free Will*; Clarke, "Toward a Credible Agent-Causation Account of Free Will;" and O'Connor, "Agent Causation."

Pass the Salt

Imran has good manners. He says 'please' and 'thank you', and he nearly always responds appropriately to polite requests from others. One evening at dinner, Yasmine says to him, "Pass the salt, please." In response, Imran exercises the power of agent causation and causes himself to pass the salt.

Let us call Imran's action of passing the salt "e_5". So we have a situation that can be illustrated by Figure 1.

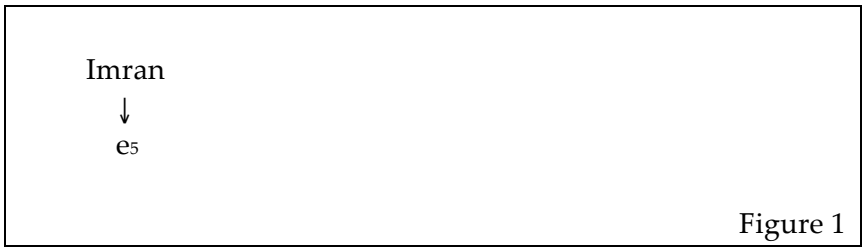

Figure 1

The arrow in Figure 1 from Imran to e_5 indicates that there is a causal relation between these two items. That is, the arrow indicates that Imran causes e_5. The agent causation theorist no doubt owes us an account of what exactly agent causation is, and how it works, and perhaps when it occurs. But if we set those questions aside for now, it is easy to appreciate the intuitive appeal of the idea behind the theory of agent causation. If our situation really is as we have described it, with this causal relation holding between Imran and his action, then it seems clear that Imran is indeed responsible for his action, and is therefore acting freely.

Now, it is natural to think that e_5 will itself have many effects (including, for example, Yasmine's saying "Thank you"). So e_5 is an event that is caused by an agent, and that itself causes many subsequent events. But it is also natural to think that e_5, in addition to being caused by its agent, is also caused by certain events. There is, for example, Yasmine's saying "Pass the salt, please." Call that event "e_4". I think it's clear that, on any plausible account of causation between events, it will turn out that e_4 is a cause of e_5. Take for example the counterfactual account of event causation.[3] It is certainly true that there is the right kind of counterfactual dependence between e_4 and e_5: if e_4

[3] See David Lewis, "Causation."

had not happened, then e_5 would not have occurred. So on the counterfactual account of event causation it will turn out that e_4 causes e_5. Or take the causes-as-probability-raisers-of-processes account of event causation.[4] It is uncontroversially the case that Yasmine's request (e_4) raised the probability of the process (involving messages going out from Imran's brain to his muscles) that resulted in Imran's passing the salt (e_5). In fact, I would go so far as to say that it would be downright bizarre to deny that e_4 causes e_5. For it would be bizarre to say that, although Yasmine asked Imran to pass the salt, and although he then passed the salt, his passing the salt was not caused by her request.

This means that we can add to our original diagram of the situation, taking into account both that e_5 causes subsequent events and that e_5 is itself caused by e_4. This new information is incorporated into Figure 2.

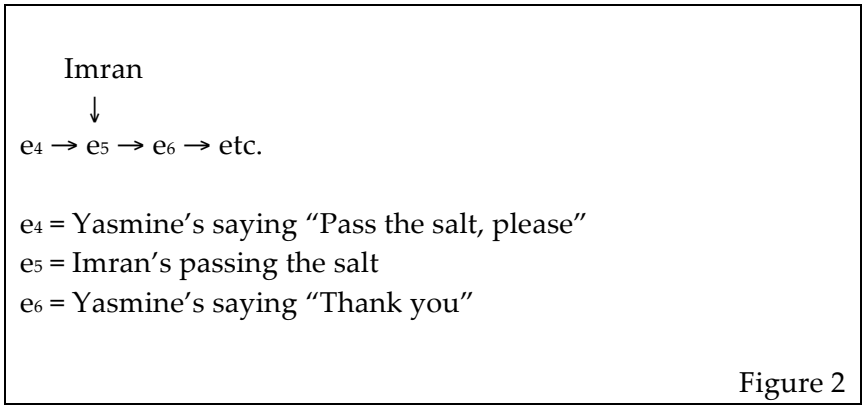

Figure 2

As Figure 2 illustrates, we have here a case in which two different entities (e_4 and Imran) are both causes of a single event (e_5). This is a common enough phenomenon, although in many instances of the phenomenon, the two causes in question are both events. In general, there are two main ways in which it can happen that two events both cause a third event. These two ways are illustrated by the following two diagrams.

---

[4] See Jonathan Schaffer, "Causes as Probability Raisers of Processes."

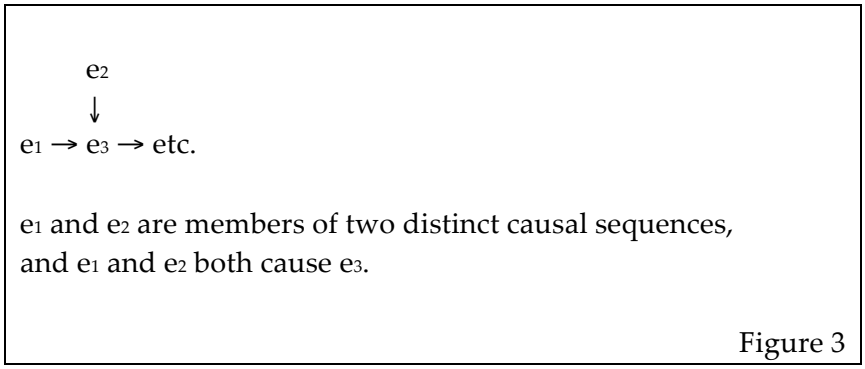

Figure 3

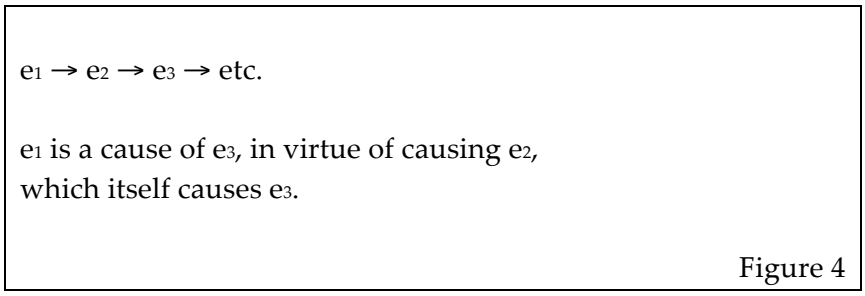

Figure 4

In Figure 3 we have two distinct causal sequences leading to a single effect (namely, e_3). One of these sequences contains e_1 but does not contain e_2, while the other causal sequence contains e_2 but not e_1. In Figure 4, on the other hand, we have just one causal sequence leading to e_3, and this one causal sequence contains both e_1 and e_2.

There is an interesting question about the way in which e_4 and Imran both cause e_5. Is it that e_4 and Imran are both causes of e_5 in the way that e_1 and e_2 are both causes of e_3 in Figure 3, i.e., as two causes that are contained in distinct causal sequences leading to a common effect? Or is it instead that e_4 and Imran are both causes of e_5 in the way that e_1 and e_2 are both causes of e_3 in Figure 4, i.e., as two causes that are contained in a single causal sequence, with one cause further upstream in that sequence than the other?

I think the answer to this question is that Imran's case is similar in a crucial way to the one represented by Figure 3, and importantly different from the case represented by Figure 4. In order to say why, I must first say what I mean by a "causal sequence." As I am using the phrase, a causal sequence is a series of causes and effects, with later members of the series being caused by earlier members. In Figure 3, since neither one of e_1 and e_2 causes the other, e_1

and e_2 are members of distinct causal sequences. But in Figure 4, since e_1 causes e_2, they are both members of the same causal sequence.[5]

Now consider again the case of Imran and e_4. It's clear that Imran does not cause e_4 (since e_4 is Yasmine's saying "Pass the salt, please"). And it should be equally clear that e_4 does not cause Imran, for the simple reason that Imran is not the right kind of thing to be caused (for he is not an event). Thus, since Imran and e_4 both cause e_5, and since neither one of Imran and e_4 causes the other, it follows that Imran and e_4 are members of distinct causal sequences leading to e_5. And so Imran's case is similar in this respect to the case represented by Figure 3 (and importantly different from the case represented by Figure 4).

It might be wondered at this point whether we should say that there is in our story such an event as Imran's causing himself to pass the salt. After all, it's quite plausible to think that Imran's causing e_5 is a case of Imran *doing* something, which something would presumably count as an action performed by Imran and, hence, an event. So it is natural to think that our example contains an event – which we can call "$e_{4.5}$" – appropriately described as Imran's causing e_5, and, moreover, that e_4 causes e_5 in virtue of causing this other event.[6] If this is right, then the situation with respect to e_4's causation of e_5 in our example is similar in at least one respect to the situation involving e_1's causation of e_3 in Figure 4; for in each case we have the causing of a particular effect through the causing of a cause of that effect. If we do say that our story contains an event appropriately described as Imran's causing himself to pass the salt, then a more perspicuous diagram of our example would look like this.

---

[5] More perspicuously: In Figure 4, since e_1 causes e_2, which in turn causes e_3, there is a single causal sequence that contains both e_1 and e_2; whereas in the case shown in Figure 3 (we can suppose) there is no causal sequence that contains both e_1 and e_2.

[6] It might be thought that if we say there is such an event as $e_{4.5}$, then we will also have to posit an infinite series of further events, including Imran's causing $e_{4.5}$, Imran's causing that further event, and so on. I think there are two possible ways for the agent causation theorist to go here. Way 1: There is no such further event as Imran's causing $e_{4.5}$, since the only event in the example that we should take to be caused by Imran is e_5. (According to this line, then, $e_{4.5}$ is caused by e_4, but is not caused by Imran himself.) Way 2: There is such an infinite regress, but it is a benign one, since by causing e_5 Imran automatically causes $e_{4.5}$, as well as all of the other events in the series.

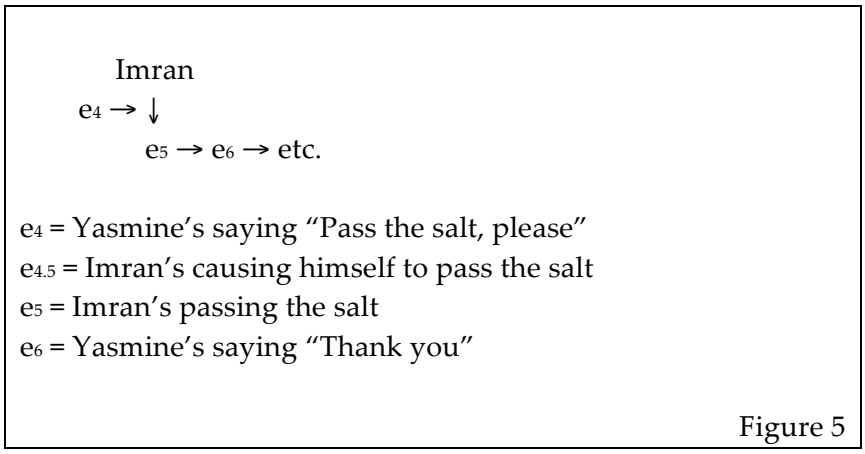

Figure 5

Notice that in Figure 5, the arrow from e_4 points, not directly at e_5, but, rather, at the arrow between Imran and e_5. This indicates that on the model currently under consideration, e_4 causes e_5 in virtue of causing something else (namely, Imran's causing himself to pass the salt). Thus Figure 5 captures the respect in which, on the current proposal, the case of Imran is similar to the example shown in Figure 4, where one event causes another event in virtue of causing an intermediate event. But notice also that Figure 5 contains a distinct causal sequence leading to e_5, namely, the one that originates with Imran himself. Thus Figure 5 also captures the way in which, even on the current proposal, our example involving Imran is similar to the one shown in Figure 3.

In any event, what our example shows is that it would be a mistake for the agent causation theorist to insist that an action can be morally free only if it is not caused by any previous event. For if anything like the theory of agent causation is true, then e_5 will be a paradigmatic case of a free action. But it is clear that e_5 is caused by e_4.

Although our example shows that the theory of agent causation should be formulated in such a way as to entail that an action can be free and yet still be *caused* by previous events, it does not by itself show that the theory of agent causation should be formulated so as to entail that an action can be free even if it is *completely determined* by previous events. For that we would need different examples, which I will not have the space to consider in this paper.[7] But I hope I have said enough here to give a clear idea of the main motivation

[7] I give such examples in Markosian, "A Compatibilist Version of the Theory of Agent Causation."

for the claim that agent causation theorists should be compatibilists. Now I will turn to my argument for the claim that compatibilists should be agent causation theorists.

2 Six Main Problems for Compatibilism

The first of the six problems facing compatibilism that I want to consider is simply the powerful intuition that you cannot be responsible for actions that were determined before you were even born. After all, you are not responsible for events that were going on before you were born, and if those events determine what you will do tomorrow, then it is admittedly difficult to see how you can be responsible for what you will do tomorrow.

The second problem facing compatibilism that I want to consider is a variation on Peter van Inwagen's modal argument involving the inference rule that he calls (β).[8] (β) concerns a special modal operator that, for our purposes, can be defined as follows (where 'p' ranges over sentences that express propositions).

> Np =df p and it has never been even partly up to any currently existing human whether p.[9]

The rule itself can be formulated like this.

(β) $N(p \supset q)$
Np
———————
Nq

---

[8] See van Inwagen, *An Essay on Free Will*, pp. 93ff.

[9] The definition of 'N' that van Inwagen gives in *An Essay on Free Will* goes like this: Np =df p and no one has, or ever had, any choice about whether p. There are a number of variations on this definition of 'N' that are interesting, and different in important ways. But in order to conserve time and space, the only definition of 'N' that I will discuss in this paper is the one given in the text. Although I won't be able to defend this claim here, I think that the points I will make regarding 'N' generalize sufficiently to apply to most other interesting definitions of the operator.

(β) certainly looks like a valid inference rule. (After all, the alethic-modality-analogue of (β), with a box in place of 'N', is valid in S5, the most popular system of modal logic.) But here is an application of (β). (Let P_0 = a sentence expressing the state of the world at some time in the remote past, let L = a sentence expressing the laws of nature, and let A = a sentence describing some action that will be performed by you tomorrow.)

An Application of (β)

(1) $N((P_0 \,\&\, L) \supset A)$
(2) $N(P_0 \,\&\, L)$
———————
(3) NA

Premise (1) of this inference says that the conjunction of P_0 and L implies A, and that it has never been up to anyone whether this implication holds. It is hard to imagine anyone denying that this follows from the assumption that determinism is true. Premise (2) of the inference says that the conjunction of P_0 and L is true, and that it has never been up to anyone whether this is so. Certain compatibilists – "altered-past compatibilists" and "altered-law compatibilists" – have denied this, but doing so seems fairly implausible to most people.[10] Since the compatibilist wants to deny (3), then, it looks like her best hope is to reject (β), which sanctions the inference from (1) and (2) to (3).

The third problem for compatibilism that I want to consider is what I will call the ability-to-do-otherwise objection. It's natural to think that the ability to do otherwise is a necessary condition for freedom and responsibility. For example, it's natural to think that if you freely rob a bank, so that you are responsible for your action, then it must be true that you could have refrained from robbing the bank. But it is also natural to think that determinism entails that no one ever has the ability to do anything other than whatever he or she actually does. So it is natural to think that determinism entails that no one ever has either moral freedom or moral responsibility.

---

[10] On both kinds of compatibilism, see John Martin Fischer, "Incompatibilism." For more on altered-law compatibilism, see also Lewis, "Are We Free to Break the Laws?"

The fourth problem for compatibilism involves manipulation arguments.[11] The basic strategy behind these arguments is to describe a case in which an action by some agent, S, satisfies the compatibilist's favorite conditions for freedom, even though S is in fact being manipulated by some powerful being in such a way that it is extremely intuitive to say that S's action should not count as a free action. Consider the following example.

> **Tom and Jerry**
>
> Tom is a good person who would never steal money from the local food bank, even if he had a golden opportunity to do so. Jerry is a bad person who would. Both are paradigm cases of free agents, according to the compatibilist, who act according to their beliefs and desires and personalities, without any unusual forces (like strings or hypnosis) constraining their behavior. But one night some powerful alien neuroscientists, using sophisticated laser-surgery techniques, alter Tom's brain in such a way that, after the surgery, he is relevantly like Jerry. That is, after the surgery, Tom *is* the kind of person who would (acting according to his own beliefs, desires, and personality) steal money from the local food bank. The next day Tom, like Jerry, is presented with a golden opportunity to steal from his local food bank, and Tom, like Jerry, does indeed steal the money.

Since Jerry is a paradigmatic case of a free agent, according to the compatibilist, she must say that Jerry is acting freely when he steals the money. And since Tom is relevantly like Jerry (and so is acting from his beliefs, etc.), it appears that the compatibilist must also say that Tom is acting freely when he steals the money from his local food bank. But most of us will want to say that poor Tom is not acting freely, due to the manipulation by the aliens.

---

[11] See for example Taylor, *Metaphysics*, pp. 43-44; Kane, *The Significance of Free Will*, pp. 65-71; Pereboom, *Living Without Free Will*, Ch. 4; and Mele, *Free Will and Luck*, pp. 188-195.

The fifth problem for compatibilism is the problem of constitutive luck. Here is how it's described by Dana K. Nelkin in *The Stanford Encyclopedia of Philosophy*.

> Constitutive luck is luck in who one is, or in the traits and dispositions that one has. Since our genes, care-givers, peers, and other environmental influences all contribute to making us who we are (and since we have no control over these) it seems that who we are is at least largely a matter of luck. Since how we act is partly a function of who we are, the existence of constitutive luck entails that what actions we perform depends on luck, too.[12]

The sixth and final problem for compatibilism that I want to consider here is one that might be called "the fairness objection." It is sometimes said that the negative attitudes normally associated with blame (such as indignation, condemnation, and resentment) would not really be appropriate, or fair, if determinism were true.[13] *Tout comprendre c'est tout pardonner*, and all that. Yet we do get indignant; we do condemn and resent when others perform morally wrong actions. We also feel shame and humiliation when we do the wrong thing ourselves. How are any of these attitudes fair, if all of the actions in question are pre-determined by conditions outside of their agents?

3 COMTAC Solutions to the Problems for Compatibilism

I take it that the standard response to our first problem (the powerful intuition that we can't be responsible for actions that were somehow determined before we were born) goes something like this. When you perform an action, as long as your action is caused by a decision within you, which itself is caused by your beliefs and desires, which are the results of your personality, etc., then you are responsible for that action, even if the action was determined by events going on before you were born. Sometimes the point is put this way:

---

[12] Nelkin, "Moral Luck."

[13] I'm grateful to an anonymous referee for encouraging me to address this objection, as well as the previous two.

It's okay that there is a long causal sequence leading from outside of you to your action, as long as this sequence "flows through you" in the right way.

Here is why this response often fails to convince people. Even if it's true that some action of yours is caused by your choices, beliefs, and desires, and also true that the relevant causal sequence flows through you in some appropriate way, it remains true that those choices, beliefs, and desires, not to mention the whole causal sequence (wherever it flows), are all caused by events outside of you that were going on before you were born. (If determinism is true, that is.) In other words, many people find it hard to see how the fact that the causal sequence flows through you is enough to make you responsible for your action.

The debate on this matter between compatibilists and incompatibilists tends to proceed like this. The compatibilists introduce a fancy definition of 'free action' in response to the strong incompatibilist intuition. (This is the part where the compatibilists characterize the right way for the causal sequence to flow through the agent.) Then the incompatibilists devise apparent counterexamples to that definition, in which the causal sequence flows in the prescribed way, but the action in question doesn't seem free; whereupon the compatibilists introduce new definitions in response to those apparent counterexamples. As the cycle repeats itself, with more and more definitions and apparent counterexamples, the compatibilist definitions and views become more and more complicated. And, I submit, the more complicated the compatibilist views become in response to the apparent counterexamples, the less strength they have against the original intuition that you can't be responsible for actions that were determined before you were born.

Now, I suggest that what the compatibilist needs to counter this simple yet powerful intuition is a simple but powerful reason for saying that your actions are free and that you are responsible for them, even though they are caused by events that occurred long before you were born. And I can't imagine a simpler or stronger reason than this: you yourself are a cause of your own actions. Since you caused them, you are responsible for them. And since moral freedom is a prerequisite for moral responsibility, you are acting freely.

As an illustration of how powerful an antidote the idea of agent causation is to the incompatibilist's intuition, consider again our example from above, in which Imran's passing the salt is caused both by Imran and by Yasmine's

request. But now let's stipulate that Yasmine's request is itself caused by events going on a million years ago. So here's the thought experiment: ask yourself whether, given all of these assumptions, it is plausible to say that Imran is responsible for his action. And here's a revised version of the diagram to help you think about the case.

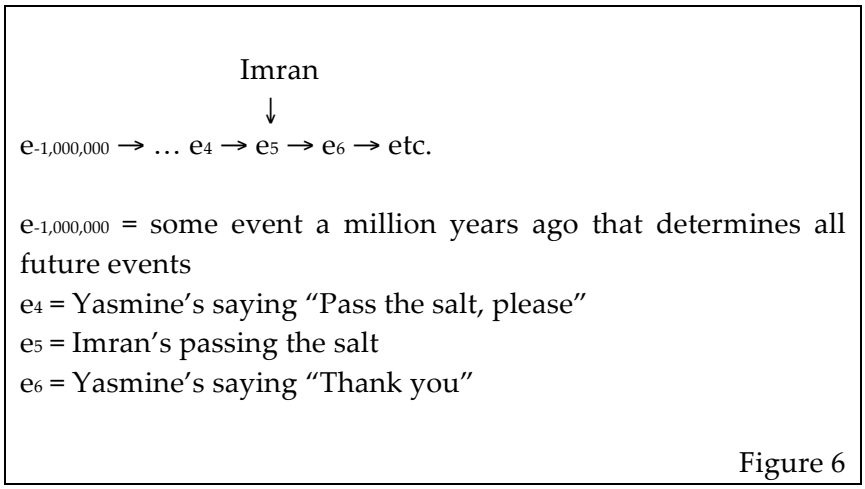

Figure 6

Notice that in this case there are two distinct causal sequences leading to e_5. One is long, and goes all the way back to $e_{-1,000,000}$. But the other one is very short; it includes only Imran and e_5.

I maintain that it is extremely plausible to say that Imran is responsible for his action in this case. After all, he himself causes it! Also, there's a (very short) causal sequence that he initiates and that contains the action in question. If you proceed backwards along that causal sequence, looking for whatever entity is responsible, the end of your backward tracing will be none other than Imran himself. Put another way: as you move back along this particular causal sequence, Imran can truthfully say, "The buck stops here." And he can also truthfully say, "That causal sequence you're investigating: it begins right here with me."

Here is what I take to be the upshot of this little thought experiment: the incompatibilist intuition that a person is not responsible for an action that is caused by external events – including events that occurred before he was born – virtually *evaporates* when we add to our description of the case the claim that the agent himself is a cause of his own action. And here is my explanation for why this happens. In general, we tend to forget about the phenomenon of

multiple causal sequences leading to a single event (and the multiple chains of responsibility that can accompany this phenomenon). So when we first hear of a case in which an event is caused by some series of other events, we tend to think that that is the whole causal story. And hence we tend to think that nothing else, apart from those other events, is responsible for the effect in question. But when we focus on a case like Imran's, explicitly involving multiple causal sequences, we quickly remember that an event can have two distinct causes, each of which is in some way responsible for the effect. And so we quickly retract our claim that nothing else (besides the relevant series of events leading up to the event in question) can be responsible for that event. (It is worth noting here that whatever was originally plausible in the compatibilist's fancy definitions of freedom (e.g., the idea that free actions are the results of choices, beliefs, desires, etc.) can be incorporated into COMTAC. For it is very natural to think that a full-blown account of agent causation will include the principle that any action that is caused by you is also the result of your choices, beliefs, etc.)

So much for our first problem facing compatibilism. Now I turn to the second one: van Inwagen's argument from (β). I don't know what deserves to be called the standard compatibilist response to this problem, but here is what strikes me as the most promising one. The standard compatibilist (i.e., one who does not believe in agent causation) can say that (β) is invalid because it *can* be at least partly up to you what you will do next, even if it was not up to you either what was going on a million years ago or whether what was going on a million years ago determined what you will do next. The idea being that all that's required in order for it to be partly up to you whether you do something is that the causal sequence leading up to your doing that thing flows through you in the right way.

How convincing is this as a response to van Inwagen's modal argument? This is a complicated and deep question, and I don't have the space here to consider the issue properly. But I will say this: although making this kind of objection to (β)-type rules strikes me as the best of the responses available to the standard compatibilist, I don't find the response to be all that convincing. For it seems strange to say that, just because the causal sequence leading up to your action involves choices, beliefs, and desires within you, it is somehow partly up to you what you will do. Since, after all, those choices, beliefs, and desires, like your action itself, are all determined by events that occurred a million years ago, and that were not at all up to you.

Now let me sketch a variation on this response to the modal argument that is available to the COMTACer. The COMTACer can say that (β) is invalid precisely because of cases involving agent causation. Here is the idea. Suppose it was not up to you either what was going on a million years ago, or whether what was going on a million years ago determined what you will do next. Still, if we assume that you have the power of agent causation, and that you will literally be a cause of your next action, then it's very plausible (I think) to say that it *is* partly up to you what you will do next. After all, on this assumption, there is a causal sequence that is initiated by you and that leads to your next action. Hence it is partly up to you what you will do next in the sense that you personally are among the causes of what you will do next.

I think this point generalizes. There are other versions of (β), corresponding to other ways of understanding the operator 'N'. I hope I have said enough to make it pretty clear that similar remarks will apply to many of these other versions of (β). The upshot, I claim, is that the compatibilist who also believes in agent causation has a much stronger reason for rejecting (β)-type rules, and hence has a much more plausible response to van Inwagen's modal argument, than the ordinary compatibilist.[14]

Our third problem for compatibilism was the ability-to-do-otherwise objection. There are two main standard compatibilist responses to this objection. The first involves admitting that the ability to do otherwise is a necessary condition for freedom and responsibility, but claiming that on the correct analysis of ability, namely, the contextualist analysis, the ability to do otherwise is compatible with determinism. According to the contextualist analysis of ability, to say that S is *able* to do A is to say that S's doing A is consistent with the relevant facts, where which facts are relevant is determined by features of the context of utterance.[15]

The second of the standard compatibilist responses to the ability-to-do-otherwise objection involves saying that the ability to do otherwise is not a

---

[14] As Joseph Keim Campbell pointed out in his comments on the APA version of this paper, all the compatibilist really needs to appeal to, in order to gain the relevant advantage (with respect to (β)-type rules) over standard compatibilism, is the *possibility* of agent causation. But it seems to me that anyone who thinks agent causation is possible should also believe it to be something that we actually do.

[15] The contextualist analysis of ability is suggested by Lewis in his "The Paradoxes of Time Travel."

necessary condition for freedom and responsibility. Typically, the compatibilist who takes this line appeals to Frankfurt examples to bolster the claim.[16]

Although I endorse the contextualist analysis of ability, I think that it is a mistake for the compatibilist to respond to the ability-to-do-otherwise objection by appealing to that account of ability. The reason is that there is a problem for this compatibilist response that has not been adequately noted in the literature. On the contextualist account of ability, whether someone is able to do a certain thing can vary from context to context and is, in general, a relativistic matter. Which means that if we make the ability to do otherwise a necessary condition for moral freedom and responsibility, and adopt the contextualist account of ability, then we will have to say that whether a given action is done freely, and whether its agent is responsible for it, are relativistic matters that can vary from context to context. But this is an unacceptable result. For we don't want to allow that when person A says "C acted freely and is morally responsible for her action," and person B says "C did not act freely and is not morally responsible for her action," A and B both speak the truth. For this reason, I think the first response to the ability-to-otherwise objection is pretty clearly untenable.

So I think that the compatibilist's best response to the ability-to-do-otherwise objection is the second response (denying that the ability to do otherwise is a necessary condition for moral responsibility). Moreover, I think that appealing to agent causation can help out the compatibilist who opts for this response to the objection. Here's why. It's very hard to convince people, just by appealing to Frankfurt examples, that the ability to do otherwise is not a necessary condition for moral responsibility. For people tend to think that the examples are weird, or somehow don't count; or else they resort to a fine-grained individuation of events in order to insist that the Frankfurt examples don't show what they are supposed to show. But it's much easier to convince people that the ability to do otherwise is not a necessary condition for moral responsibility if you appeal to agent causation.

In order to see this, consider again our Pass the Salt example. Suppose, as before, that Imran's passing the salt is determined by Yasmine's request. And suppose further that we manage to get ourselves into a context in which that

---

[16] On Frankfurt examples, see Harry Frankfurt, "Alternate Possibilities and Moral Responsibility."

very fact is among the relevant facts. Then, even on the contextualist analysis of ability, we will speak truly when we say that Imran is not able to do anything besides pass the salt. But now add to the story that Imran exercises the power of agent causation, and causes himself to pass the salt. And ask yourself whether we should say that Imran is responsible for what he does in this case.

I think that the obvious answer is *Yes*. Even if Imran is such a polite guy that he couldn't do anything else besides pass the salt when Yasmine asked him, still, he did cause his own action. And it would be a shame to penalize Imran for being such a polite guy, as we would be doing if we said that he is not responsible, but that a less polite person in his shoes, who is able to do otherwise, is responsible. So it is very plausible to say that Imran is responsible for his action, even though he was not able to do otherwise.

Our fourth problem for compatibilism involved manipulation arguments, like the one suggested above by the case of Tom and Jerry. It is tempting for the compatibilist to respond to such an argument by insisting that Tom is not in the same boat as Jerry precisely because his brain – and hence his personality – was manipulated by the aliens. But in fact it is not so easy to make this response work.[17] The problem is that we are all manipulated by other agents in a variety of different ways, so that it is not at all clear that there is any principled place to draw the line between manipulation that results in a free agent (such as the moral training that we inflict upon our children) and manipulation that results in an agent who is not free (like the case of poor Tom). For the two extremes on this spectrum can be connected by a series of cases in such a way that adjacent members of the series are extremely similar to one another in all relevant respects.

Here is how an appeal to agent causation can help the compatibilist with this problem. Suppose the compatibilist endorses COMTAC. Then she can point out that there are two possible ways of filling out the details of Tom's story. Either the aliens alter Tom's brain in such a way that the resulting person causes his own action of stealing, or not. If the aliens succeed in bringing it about that Tom causes his own action when he steals, then it is no longer implausible to say that Tom is morally responsible for his action (and hence is acting freely). For in that case, Tom is a cause of his action. We will no

---

[17] What follows is a variation on the four-case argument developed by Derk Pereboom. See Pereboom, *Living Without Free Will*, Ch. 4.

doubt also want to criticize the aliens for changing Tom from an honest person into a thief, but we can at the same time say that the resulting individual is a bad person who performs a morally wrong action. And although a standard compatibilist can say some similar things without appealing to the idea of agent causation, she (the standard compatibilist) cannot say the one thing that makes this response on the part of the proponent of COMTAC especially powerful: that even though there is a causal chain (involving the manipulative aliens) leading from outside of Tom to his action, there is also another causal chain leading to Tom's action *that originates with Tom himself*, and this is the reason why he is responsible for what he does.

I mentioned above that there are two possible ways of filling out the details of Tom's story. So far we have seen that if the story is told in such a way that Tom causes his own act of stealing, then it is quite plausible to say that he is therefore responsible for what he does, despite the manipulation by the aliens. If, on the other hand, the details of the story are filled out in such a way that the aliens do not succeed in bringing it about that Tom causes his own action when he steals, then the COMTACer will have an even easier time of dealing with the example. For in that case the COMTACer will of course say that Tom is not responsible for stealing the money from the food bank, while Jerry is, because of a crucial difference between Tom and Jerry: Jerry causes his own action of stealing, but Tom does not cause his own theft. And this does indeed seem like the kind of principled reason that could support such a distinction in moral status between the two actions.

The fifth of our problems for compatibilism was the problem of constitutive luck. If you are not responsible for certain aspects of your nature, and if those aspects of your nature result in your actions, then how can you be responsible for your actions? The standard compatibilist will no doubt say something like the following in response to this problem.

> It's true that there is such a thing as constitutive luck. But this does not undermine moral responsibility. If your actions flow from your character in the right way, then you are responsible for them. If not, then not. The fact that it is partly (or even largely) a matter of luck what your character is like is neither here nor there.

The problem with this response is that its proponent is forced to bite a rather large bullet: she must say that there is this crucial factor – your character –

that must cause your actions in order for you to be responsible for those actions, but she must also admit that this crucial cause of your actions is itself caused by other factors that are completely outside of your control.

By now it should be clear that the proponent of COMTAC has something much more plausible to say here. For the COMTACer says that the crucial factor is whether you are or are not a cause of your own actions. Which means that he also gets to say two further things: (1) when the crucial factor is present, there is a causal sequence leading to your action that originates with you; and (2) when the crucial factor is present, the relevant cause of your action (the one in virtue of which you are acting freely) is not itself caused by other factors that are completely outside of your control – for the relevant cause on his view is an agent, and (as we have noted above) an agent is simply not the right kind of thing to be caused.

Our sixth and final problem for compatibilism involved the fairness objection. How can it be fair to have the negative attitudes that are normally associated with blame (such as condemnation) if determinism is true? The standard compatibilist will have to say that we condemn wrongdoers because their personalities are such that they are disposed to do the wrong things that they do, and that this is fair even though the causes of the wrongdoers' personalities are outside of them. But to many neutral parties this response will not seem very compelling. Insofar as wrongdoers (like everyone and everything in the world, if determinism is true) are merely caught up in causal chains that go back millions of years, it is hard to see what is fair about holding such negative attitudes toward them.

Matters are very different, however, if the compatibilist is free to appeal to the idea of agent causation. For then he can claim that, in addition to all of the million-year-old causal chains leading up to any action by any human agent, there will also be, for each such action, a very short causal chain leading to that action – a causal chain that does not extend outside of the agent, and that in fact originates with that very agent. And if all of that is true, then the feeling that it is unfair to condemn someone for doing the wrong thing, like the original incompatibilist's intuition that you cannot be responsible for actions that were determined before you were even born, seems to evaporate into thin air. It seems eminently fair to condemn a wrongdoer, and to be indignant and even resentful, if that wrongdoer has *caused herself* to perform the wrongful action in question.

4 Conclusion

I'm playing the role of matchmaker here. The match I am attempting to arrange is between the agent causation theorist and the compatibilist. In my earlier paper, I tried to carry out the first part of my matchmaker's project – convincing the agent causation theorist to be a compatibilist. That was the easier of the two parts, because the agent causation theorist already believes in agent causation, and because it was relatively easy to show that there are major problems with each of the incompatibilist versions of the theory of agent causation. Thus it was relatively easy to show that the agent causation theorist should give up incompatibilism and embrace COMTAC.

In this paper I have tried to carry out the second part of my matchmaker's project – convincing the compatibilist to be an agent causation theorist. This, it seems to me, is a much harder sell, for one main reason: the assumption that there is such a thing as agent causation is a radical assumption. It requires believing in a kind of causation that involves, not events, but *things* (namely, *agents*) as the first of the two relata in causal relations. And it involves saying that these agents have an amazing power: the power to initiate causal sequences.

But selling compatibilists on the idea of COMTAC is also a hard sell for another reason: the necessary assumption – that there is such a thing as agent causation – is one that leads to many further questions that I have not addressed, including the following:[18]

(Q1) Even if we assume that there really is such a phenomenon as agent causation, can we say more about what exactly it is and how it is supposed to work?

(Q2) What kinds of things can have the power of agent causation?

(Q3) Is there any reason at all to think that we humans actually have the power of agent causation?

(Q4) Are cases of agent causation always cases in which an agent and some event are both causes of the action in question, the

---

[18] For some discussion of questions like at least some of the following, see Clarke, "Toward a Credible Agent-Causation Account of Free Will" and *Libertarian Accounts of Free Will*; and O'Connor, "Agent Causation" and *Persons and Causes*.

way Imran himself and Yasmine's request are both causes of Imran's action in our Pass the Salt example? Or do cases of agent causation sometimes involve the agent causing an action that is not also caused by previous events?

(Q5) Does agent causation supervene on event causation, in the sense that any two worlds that differ with respect to the facts about agent causation must also differ with respect to the facts about event causation?

(Q6) Do agent-caused actions typically depend counterfactually on their being caused by their agents, so that if the agent hadn't caused the action then the action would not have occurred?

(Q7) How much of a robust realist about agent causation must one be in order to reap the benefits for compatibilism extolled in this paper? In particular, must one say that agent causation is an irreducible, fundamental feature of the world in order to enjoy those benefits, or can one get away with holding merely that agent causation is a feature of the world, but one that is somehow reducible and non-fundamental?

Although I'm fairly optimistic about the prospects of persuading the agent causation theorist to be a compatibilist, I am much less sanguine about the chances of convincing the compatibilist to be an agent causation theorist, and I certainly don't imagine that I've closed the deal yet. What I do hope to have accomplished in this paper is a more modest goal. I hope to have convinced the compatibilist that the benefits associated with adopting the theory of agent causation are sufficiently great that it is worth at least exploring the costs. That is, I hope to have convinced the compatibilist that the theory of agent causation is attractive enough as a potential partner that it is at least worth looking into questions like these seven.[19]

---

[19] Earlier versions of this paper were presented at Western Washington University, the 2001 Inland Northwest Philosophy Conference, the 2002 Meeting of the Pacific Division of the American Philosophical Association, Macquarie University, the University of Durham, the 2005 Bled Philosophy Conference, the Australian National University, and Vancouver Island University. I am indebted to members of all eight audiences for helpful criticism. I am also grateful to Andrew Egan, Kris McDaniel,

Works Referred To

Chisholm, Roderick, "Human Freedom and the Self," presented as the Lindley Lecture at the University of Kansas, 1964 (reprinted in Chisholm, Roderick, *On Metaphysics* (Minneapolis: University of Minnesota Press, 1989)).

Clarke, Randolph, "Toward a Credible Agent-Causation Account of Free Will," in O'Connor, Timothy (ed.), *Agents, Causes, and Events* (Oxford: Oxford University Press, 1995).

Clarke, Randolph, *Libertarian Accounts of Free Will* (Oxford: Oxford University Press, 2003).

Fischer, John Martin, "Incompatibilism," *Philosophical Studies* **43** (1983), pp. 127-137.

Frankfurt, Harry, "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy* **66** (1969), pp. 829-839.

Kane, Robert, *The Significance of Free Will* (New York: Oxford University Press, 1996).

Lewis, David, "Are We Free to Break the Laws?" *Theoria* **47** (1981), pp. 113-121.

Lewis, David, "Causation," in his *Philosophical Papers*, Vol. II (New York: Oxford University Press, 1986), pp. 172-213.

Lewis, David "The Paradoxes of Time Travel," in David Lewis, *Philosophical Papers, Volume II* (Oxford: Oxford University Press, 1986), pp. 67-80.

Markosian, Ned, "A Compatibilist Version of the Theory of Agent Causation," *Pacific Philosophical Quarterly* **80** (1999), pp. 257-277.

Sarah McGrath, Ted Sider, Ryan Wasserman, Brian Weatherson, and an anonymous referee for helpful comments on earlier versions of the paper.

Mele, Alfred R., *Free Will and Luck* (Oxford: Oxford University Press, 2006).

Nelkin, Dana K., "Moral Luck," *The Stanford Encyclopedia of Philosophy* (2008).

O'Connor, Timothy, "Agent Causation," in O'Connor, Timothy (ed.), *Agents, Causes, and Events* (Oxford: Oxford University Press, 1995).

O'Connor, Timothy, *Persons and Causes* (New York: Oxford University Press, 2000).

Pereboom, Derk, *Living Without Free Will* (Cambridge University Press, 2001).

Schaffer, Jonathan, "Causes as Probability Raisers of Processes," *Journal of Philosophy* **98** (2001), pp. 75-92.

Suarez, Francisco, *Disputationes Metaphysicae*.

Taylor, Richard, *Action and Purpose* (Englewood Cliffs, NJ: Prentice-Hall, 1966).

Taylor, *Metaphysics, 4th Edition* (Englewood Cliffs, NJ: Prentice-Hall, 1974).

Turner, Jason, and Nahmias, Eddy, "Are the Folk Agent-Causationists?" *Mind & Language* **21** (2006), pp. 597-609.

Van Inwagen, Peter, *An Essay on Free Will*, (Oxford: Oxford University Press, 1983).