A version of this paper appears in *Pacific Philosophical Quarterly* **80** (1999), pp. 257-277.

A COMPATIBILIST VERSION OF THE THEORY OF AGENT CAUSATION

Ned Markosian

1 Introduction

The problem of freedom and determinism has vexed philosophers for several millennia, and continues to be a topic of lively debate today. One of the proposed solutions to the problem that has received a great deal of attention is the Theory of Agent Causation. While the theory has enjoyed its share of advocates, and perhaps more than its share of critics, the theory's advocates and critics have always agreed on one thing: the Theory of Agent Causation is an incompatibilist theory. That is, both believers and nonbelievers in the theory have taken it for granted that the most plausible version of the Theory of Agent Causation is one according to which freedom and determinism are incompatible. In fact, so entrenched is this assumption that no one on either side of the debate has ever questioned it. Yet it turns out that this assumption is wrong – the most plausible version of the Theory of Agent Causation is a compatibilist one.

Or so I will argue in this paper. I begin by spelling out some terminology that will be useful in discussing the relevant issues. Then I formulate the traditional version of the Theory of Agent Causation, and consider a series of objections to it and related views. With each objection comes a corresponding revision of the theory that is motivated by that objection, and with each revision the theory becomes increasingly compatibilistic until, finally, we arrive at a completely compatibilistic version of the Theory of Agent Causation, which I take to be the most plausible version of that theory.

It is worth noting explicitly at the outset that I will not be defending the Theory of Agent Causation in this paper. Instead, I will argue for the following thesis: If we assume that there is such a thing as agent causation, and that it is relevant to freedom and responsibility, then the version of the Theory of Agent Causation that we ought to consider most promising is a

wholly compatibilistic version of that theory. If this thesis is correct, then it is a surprising truth, for until now every theorist (pro or con) who has ever written about the Theory of Agent Causation has believed just the opposite.

2 The Traditional Version of the Theory of Agent Causation

In order to discuss our topic, we'll need to settle on some terminology, as well as formulations of certain relevant views. Let's begin by agreeing that there is some sense in which an action must be free in order for its agent to be morally responsible for it.[1] And let's reserve the technical term *morally free* to describe actions that are free in the sense required for moral responsibility.[2,3] Then we can characterize the Free Will Thesis as the claim that some actions are morally free, and compatibilism as the claim that the Free Will Thesis is

---

[1] Here are some examples of what I take to be actions: *Saleem's opening the door at t1*, *Lina's signing the letter at t2*, *Till's throwing the ball at t3*.

[2] If we want to talk about morally free agents, we can simply say that an agent is morally free whenever he or she performs a morally free action.

[3] Notice that the above definition of 'morally free' does not entail that there is a connection between an action's being morally free and its agent's being able to do otherwise. That is, the above definition of 'morally free' leaves open the question of whether moral responsibility and moral freedom require the ability to do otherwise. For this reason, "freedom" in my sense may be very different from "freedom" in some other popular senses of the word.

Notice, also, that on the above definition an agent's being morally responsible for an action entails that that action is morally free, but an action's being morally free does not entail that its agent is morally responsible for that action. That is, the definition has the consequence that an action's being morally free is a necessary condition for its agent's being morally responsible for it, but does not imply one way or the other whether an action's being morally free is a sufficient condition for its agent's being morally responsible. There may be other conditions that are also required for moral responsibility. For example, it might be that the actions of a lion, say, or a young child are morally free even though the lion or the child is not morally responsible for them, in virtue of not being a rational, deliberating agent with a grasp of moral concepts like rightness and wrongness. I won't have anything to say in this paper about what must be added to an action's being morally free in order to get a set of necessary and sufficient conditions for its agent's being morally responsible for that action.

compatible with determinism. As we shall see, however, there are different versions of determinism, some weaker and some stronger. Thus there are also different grades of compatibilism, depending on which version of determinism is said to be compatible with the Free Will Thesis. My claim is that the Theory of Agent Causation should be formulated so that it entails that the Free Will Thesis is compatible with even the strongest version of determinism. But let's begin with the traditional way of formulating the Theory of Agent Causation.

Consider the following deterministic thesis.

> **The Principle of Universal Causation:** Every event that occurs is caused by previous events.

If the Principle of Universal Causation were true, then any action you performed would have causes going back to before you ever existed.[4] It's hard to see how you could be responsible for any of your actions in that case. For it seems clear that you cannot be responsible for events that were going on before you ever existed. And if such events cause some action that you perform today, then it is hard to see how you could be responsible for that action. Meanwhile, it's not clear that simply denying the Principle of

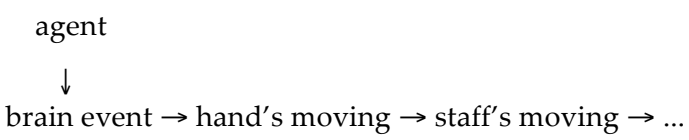
---

[4] This claim needs to qualified. For one thing, it rests on the assumption that you haven't always existed. (But considering the intended audience of this paper, that seems like a fair assumption.) For another thing, the claim requires a second assumption, having to do with the possibility of an infinite series of events such that (i) the series has a last member but no first member, and (ii) for some specific time, t, earlier and earlier events in the series approach but never reach t. (Here is one way there could be such a series: it could be that each event in the series (except the last) has one-half the temporal duration of its successor, so that as we move backwards through the series we find increasingly shorter and shorter events that get closer and closer to some specific time, without ever reaching that time.) Let's call such a series of events a "Zenonian" series. Then the claim that the Principle of Universal Causation entails that any action you perform will have causes going back to before you ever existed rests on the assumption that there are no such Zenonian series of events all involving you and all taking place within your lifetime.

I have no idea whether this assumption is true. But nothing of importance in what follows turns on either the assumption that there are no such Zenonian series or the claim that the Principle of Universal Causation entails that any action you perform will have causes going back to before you ever existed.

Universal Causation will allow us to say, with any plausibility, that you are sometimes responsible for your actions. For it's hard to see how you could be responsible for an action that just happens spontaneously, i.e., without any cause; and it is similarly hard to see how you could be responsible for an action that is caused by some spontaneous event within you, since it is difficult to see how any person could be responsible for any spontaneous event.[5]

The solution to this difficulty, according to the Theory of Agent Causation, is this: For an action of yours to be morally free, you have to be the cause of your own action. Thus, the Theory of Agent Causation is based on a very important – and controversial – assumption, namely, that *events* are not the only entities that can cause events. According to the Theory of Agent Causation, *agents* – things like you and me – can also sometimes cause events. Here is an example that illustrates how agent causation is supposed to work.[6]



The idea is that in this example the brain event is not caused by any previous event, but it is caused by something else, namely, the agent. The agent just somehow causes the brain event to occur, and then the brain event causes the agent's hand to move, which causes the staff he is holding to move, and so on. The further idea is that it is precisely in cases like this that our actions are morally free. For in such cases, we are the causes of our actions, and nothing else causes those actions.

---

[5] I am here barely gesturing in the direction of arguments for the relevant claims. For a detailed discussion of the main arguments concerning these matters, see van Inwagen, *An Essay on Free Will*.

[6] The example is from Aristotle by way of Chisholm. See Aristotle's *Physics*, Book VIII, Chapter 5, and Roderick Chisholm, "Human Freedom and the Self."

Here is a formulation of the traditional version of the Theory of Agent Causation, as it has been discussed by philosophers such as Aristotle, Suarez, Reid, Campbell, and Chisholm.[7]

> **TAC1:** A is *morally free* iff (i) A is caused by A's agent, and (ii) A is not caused by events outside of A's agent.

TAC1 yields the result that the action in our example is morally free. This strikes me as a very plausible result. If it is true that that agent caused his action, and also true that no events outside of the agent caused his action, then it is certainly natural to think that the agent is morally responsible for his action and, hence, that the action is free in the sense required for moral responsibility.

Now, as I said above, the Theory of Agent Causation is based on the controversial assumption that agents can cause events. Many people believe this assumption is false. Some people even think it is unintelligible. I won't discuss in this paper whether the crucial assumption is true, or even intelligible. Instead, I will assume for the purposes of the paper that the assumption is intelligible, and is in fact true. My interest is in seeing what is the best way to develop the theory that is based on this assumption.[8]

One other remark about the general notion of agent causation is in order here. A number of philosophers with whom I have discussed the notion of agent causation have initially assumed that the idea presupposes a non-materialistic conception of people. That is, a number of philosophers with whom I have discussed the notion of agent causation have initially thought that subscribing to some version of the Theory of Agent Causation requires believing that each person has a non-physical component – a Cartesian mind, or a soul. I don't think that this is right. I admit that it is a little odd at first

---

[7] See Aristotle's *Eudemian Ethics*, Book II, Chapter 6 and his *Nichomachean Ethics*, Book III. Chapters 1-5; Suarez, *Disputationes Metaphysicae*, Disputation XVIII, Section 10; Reid, *Essays on the Active Powers of Man*, Essay IV, Chapter 4; Campbell, "Is 'Free Will' a Pseudo-Problem?" and Chisholm, "Human Freedom and the Self," and *Person and Object*, Chapter 2.

[8] See Chisholm, "Human Freedom and the Self," and Reid, *Essays on the Active Powers of Man*, Essay IV, Chapter 4, for a response to the charge that the notion of agent causation is unintelligible. I personally am inclined to think that the notion is intelligible, and to respond to the charge of unintelligibility in much the way Chisholm and Reid do.

(especially for philosophers who have thought a lot about the notion of event causation, but very little (or not at all) about the notion of agent causation) to think that agents can cause events. But I don't see how assuming that a person has a mind or a soul makes the idea of agent causation any easier to understand. As I see it, if it is possible for a person with a non-physical component to cause some event within his- or herself, then it is equally possible for a person who is entirely physical to do the same thing. Thus, although I am assuming in this paper that the notion of agent causation is intelligible, I am not assuming that it presupposes a non-materialistic conception of people. As far as I can tell, the Theory of Agent Causation is consistent with materialism.

Returning now to TAC1, it should be clear how that theory is supposed to solve the traditional problem of freedom and determinism. According to TAC1, if the Principle of Universal Causation is false, and if, in particular, some actions you perform are not caused by previous events, but are caused by you, then those actions are morally free. And it should be clear that TAC1 is an incompatibilist view. For TAC1 entails that the Free Will Thesis is incompatible with the Principle of Universal Causation. I think it can be shown, however, that the Theory of Agent Causation should *not* entail that the Free Will Thesis is incompatible with the Principle of Universal Causation.

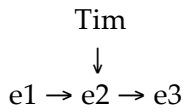
To begin with, I want to claim that it is possible for a person to be morally responsible for an event even if something else causes that event. Here is an example designed to show this.

> **Case 1:** Haley is on the seat of a dunking machine at the state fair. Kyra throws a ball at the target. Meanwhile, at the same time but independently, Manon also throws a ball at the target. The two balls hit the bull's-eye simultaneously, causing Haley to get dunked in the water.

It seems to me that Kyra is morally responsible for Haley's dunking in Case 1. Yet it remains true that an event that has nothing to do with Kyra – namely, the striking of the bull's-eye by the ball thrown by Manon – also causes Haley's dunking. For this is a case of what we might call "double causation" in which each of two independent events causes a third event. I take it that this case shows that double causation is not, in general, an impediment to moral responsibility. That is, the case shows that you can be morally responsible for an event even if something other than you causes that event. For that is how it is with Kyra (and Manon as well) in Case 1.

With this result in mind, let us now consider another example.

> **Case 2:** Tim deliberates about whether to shoot the president. Then he decides to do it, and causes the event in his brain that sends the signal to his hand telling it to pull the trigger, which it does. But at the same time, and completely independently, a rogue neuron in Tim's brain flukishly and indeterministically fires, thereby also causing the event in Tim's brain that sends the signal to his hand telling it to pull the trigger.



e1: the rogue neuron's firing

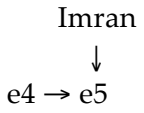
e2: the event in Tim's brain that causes his hand to pull the trigger

e3: Tim's pulling the trigger

It seems to me that if anything like the Theory of Agent Causation is true, then Tim is morally responsible for his actions (both e2 and e3) in this example, even though the firing of the rogue neuron also caused those actions. And of course, assuming that some version of the Theory of Agent Causation is true, if Tim had not caused his own actions, then he would not be morally responsible for them. But he did cause his actions, and so he is morally responsible for them.

For those who find this example to be too far-fetched, the following, more mundane example should suffice to make the same point.

> **Case 3:** Imran has good manners. He says 'please' and 'thank you', and he always responds appropriately to polite requests from others. One evening at dinner, Yasmine says to him, "Pass the salt, please." Her saying this has an effect on Imran, and as a result he exercises the power of agent causation and causes himself to pass her the salt.



e4: Yasmine's saying "Pass the salt, please"

e5: Imran's passing the salt

Imran's action in this example is caused by Imran. But it seems clear that his act of passing the salt is also caused (indirectly) by Yasmine's request. For on any plausible account of event causation, it is going to turn out (provided we tell the story in the natural way) that Yasmine's utterance causes Imran's action. To take just one example, if the counterfactual analysis of event causation is correct, then it will be true that Yasmine's utterance causes Imran's action, since there is the right kind of counterfactual dependence between these two events – if Yasmine had not asked, Imran would not have passed the salt.[9] And it should be clear that any plausible non-Humean account of event causation will also yield the same result. It's just hard to get around the fact, once you examine the situation, that Yasmine's utterance in this example had, among its various effects, Imran's passing the salt. Should we therefore say that Imran is not morally responsible for his action, since Yasmine's request caused it? I don't think so. In fact, it seems to me that if anything like the Theory of Agent Causation is true, then we should say that Imran is morally responsible for his action in this example.

Since Tim and Imran are morally responsible for their actions in Cases 2 and 3, it follows that their actions are morally free. Thus, Cases 2 and 3 show that the Theory of Agent Causation should not be formulated in a way that rules out the possibility of a morally free action's being caused by some event outside of its agent (in addition to being caused by its agent). But TAC1 rules out this possibility. Hence TAC1 is in need of repair. The upshot is that the Theory of Agent Causation should not entail that the Free Will Thesis is incompatible with the Principle of Universal Causation.

---

[9] For a presentation of the counterfactual analysis of event causation, see Lewis, "Causation."

3 A More Recent Version of the Theory of Agent Causation

Here is a deterministic thesis that is stronger than the Principle of Universal Causation.

> **Event Determinism:** Every event that occurs is made physically necessary by previous events.

It is natural to think that you can't be morally responsible for an action you perform if that action is made physically necessary by events outside of you (as opposed to being merely caused by events outside of you). For if your action is made physically necessary by events outside of you, then your action is, in a sense, completely out of your control. Here is a new version of the Theory of Agent Causation that incorporates this intuitive idea. (This version of the Theory of Agent Causation is similar to versions of the theory that have been discussed recently by Taylor, van Inwagen, Clarke, and O'Connor.[10])

> **TAC2:** A is *morally free* iff (i) A is caused by A's agent, and (ii) A is not made physically necessary by events outside of A's agent.

Since an event can be caused by previous events without being made physically necessary by those events, the Free Will Thesis is, according to TAC2, compatible with the Principle of Universal Causation. For it could be that every event is caused by previous events, but that some events – in particular, some actions we perform – are not made physically necessary by any previous events; and it could also be that some of those same actions are caused by their agents, so that they would be morally free, according to TAC2.

Although TAC2 entails that the Free Will Thesis is compatible with the Principle of Universal Causation, it also entails that the Free Will Thesis is incompatible with Event Determinism. For if Event Determinism is true, then no action will satisfy the second condition of TAC2; which means that if Event Determinism is true, then, according to TAC2, no action will be morally free.

---

[10] See Taylor, "Determinism and the Theory of Agency," *Action and Purpose*, and *Metaphysics*; van Inwagen, *An Essay on Free Will*; Clarke, "Toward a Credible Agent-Causation Account of Free Will;" and O'Connor, "Agent Causation."

But I think that the Theory of Agent Causation should not entail that the Free Will Thesis is incompatible with Event Determinism. In order to see why, consider the following example.[11]

> **Case 4:** Franny and Zoe are partners in crime. Franny, worried that Zoe will back out of their assassination plot at the last minute, places a brain-control device in Zoe's head. If Zoe decides to back out, the device will kick in and make Zoe pull the trigger anyway. When the time comes, Zoe chooses to go ahead with the crime, causing herself to do it, so that the brain-control device never has to kick in.

Notice that Franny's action (of placing the brain-control device in Zoe's head) in Case 4 makes it physically necessary that Zoe will pull the trigger. But it seems to me that Zoe is morally responsible for her action in this case. It's true that if she had not caused herself to pull the trigger, so that the brain-control device had kicked in and forced her to go through with it, then she would not have been responsible. But as it turns out, she did cause herself to pull the trigger, so she is morally responsible. This means that her action is morally free.

Case 4 shows that the Theory of Agent Causation should not be formulated in a way that rules out the possibility of a morally free action's being made physically necessary by events outside of its agent. But TAC2 rules out this possibility. Hence TAC2 is in need of repair. The upshot is that the Theory of Agent Causation should not entail that the Free Will Thesis is incompatible with Event Determinism.

4 A More Compatibilistic Version of the Theory of Agent Causation

Here is what I take to be the strongest version of Determinism:

> **Future Determinism:** There is at any instant just one physically possible future.

At any given time, according to Future Determinism, the laws of nature and the way things are at that time leave room for only one way that things could

---

[11] This is what is known in the literature as a "Frankfurt example," after Harry Frankfurt. For the original Frankfurt examples, see Frankfurt's "The Principle of Alternate Possibilities."

go from there. Thus Future Determinism does not even leave room for flexibility with regard to the presence or absence of agent causation, the way Event Determinism does. If Future Determinism is true, and if you are going to cause yourself to perform some action tomorrow, then it was physically necessary a million years ago that you were going to cause yourself to perform that action.

You might think that the Theory of Agent Causation should entail that the Free Will Thesis is at least incompatible with Future Determinism. For you might think that if someone else makes it not only physically necessary that you will perform a certain action, but also physically necessary that *you will cause yourself* to perform that action, then you shouldn't be morally responsible for that action. After all, there is a difference between such a case and Case 4 above. In Case 4, Franny did not force Zoe *to cause herself* to pull the trigger. All Franny did was force Zoe to pull the trigger. That is, Franny set things up so that, one way or another, Zoe was going to pull the trigger. But Franny still left Zoe the option of not doing so "voluntarily" – for Franny left Zoe the option of not participating in a way that involved agent causation. Which means that, according to the fundamental intuition behind the Theory of Agent Causation, Franny left Zoe the option of not participating in a way that would make her morally responsible for her action. Franny, in short, did not force Zoe to be morally responsible for a wrong action. And surely, you might think, it should not be possible for someone else to force you to perform a morally wrong action for which you would then be morally responsible.

These considerations, together with what was said above about Case 4, make the following version of the Theory of Agent Causation seem plausible.

> **TAC3:** A is *morally free* iff (i) A is caused by A's agent, and (ii) it is not physically necessary at any time before the time of A that A's agent will cause A.

TAC3 entails that the Free Will Thesis is compatible with the Principle of Universal Causation and Event Determinism, but incompatible with Future Determinism. There is, however, a problem with TAC3. In order to see the problem, consider the following example.

> **Case 5:** As Mookie grows up, he performs a series of actions that mold his own personality. Each of these actions is caused by Mookie himself, through the exercising of agent causation, and none of them is made physically necessary by events

> outside of Mookie. As a result of his own agent-caused actions, Mookie turns out to have such a fine moral character that it is physically necessary that he will always cause himself to do the right thing. One day a passerby asks Mookie for directions. Her asking makes it physically necessary that Mookie will cause himself to do the right thing and help her, which he does.

I think Mookie's action in this case is morally right. It's true that his good nature made it physically necessary that he would cause himself to help once he was asked. But it would be strange to say that his action is not a morally right action just because of this, especially since Mookie himself is the cause of his having such a good nature. (To say that Mookie's action is not a morally right action just because his good nature made it physically necessary that he would cause himself to perform that action would be in effect to penalize Mookie for causing himself to have such a good nature.) Hence Mookie's action in Case 5 is morally right. And if it is a morally right action, then he is morally responsible for his action. It follows that his action is morally free.

Case 5 shows that the Theory of Agent Causation should not be formulated in a way that rules out the possibility of its being physically necessary beforehand that an agent will cause himself to perform a morally free action. But TAC3 rules out this possibility. Hence TAC3 is in need of repair.

5 A Still More Compatibilistic Version of the Theory of Agent Causation

It is natural to think that although TAC3 goes wrong in ruling out the possibility of a person's being morally responsible in a case like Mookie's, it is nevertheless on the right track insofar as it rules out the possibility of a person's being morally responsible for performing an action made physically necessary by his or her personality in a case in which forces *outside* of the agent are the cause of the agent's having the relevant personality. Thus it is natural to think that the best way to revise TAC3 would be along the following lines. (Let 'S(A)' abbreviate 'the agent of A'.)

> **TAC4:** A is *morally free* iff both (i) A is caused by S(A); and (ii) either (a) the personality of S(A) is not such that it is physically necessary that he or she will cause A in the circumstances of A, or (b) if the personality of S(A) is such that it is physically necessary that he or she will cause A in the circumstances of A,

> then S(A) caused S(A) to have such a personality, and no events outside of S(A) made it physically necessary that he or she would do so.

Mookie's action in Case 5 would count as morally free, on this view, since is satisfies condition (ii) (in virtue of satisfying the second disjunct of that condition), and it also satisfies condition (i). But it seems to me that there is a problem even with TAC4. To see the problem, consider the following case.

> **Case 6:** Zane grows up very much like Mookie. He performs a series of actions – all caused by himself, and none made physically necessary by events outside of Zane – that shape his own personality. But in Zane's case, he does not cause himself to have a personality that makes it physically necessary that he will always cause himself to do the right thing. Instead, he causes himself to have a personality that makes it almost, but not quite, physically necessary that he will always cause himself to do the right thing. The chances of Zane's ever failing to cause himself to do the right thing, at this stage of his life, are one in a billion. Zane ends up in college, where he has a typically gifted and inspiring teacher. She teaches him so well that his personality changes. The change in his personality is really a very small one, and it occurs only because Zane is such a good student and an open-minded person. As a result of his slightly changed personality, when Zane leaves college it is literally physically necessary that he will always cause himself to do the right thing. One day soon after graduation, Zane comes upon a burning school building. He then causes himself to run inside and save hundreds of children.

Should Zane's action in Case 6 count as morally free? I don't see why not. It's true that it was the teaching of his inspiring professor that ultimately made it physically necessary – as opposed to merely physically very, very probable – that he would cause himself to rush into the building to save the children. But still, even before the teacher got to him, Zane had an amazingly upstanding personality, and was already such that there was only a one in a billion chance that he would do anything else besides cause himself to perform his heroic act. And considering what Zane was like before college, it is clear that his teacher, inspiring and marvelous though she was, did not

significantly change Zane. Surely it is implausible to think that in changing Zane as little as she did, she thereby put him in a position to be no longer morally responsible for his actions, especially since it was Zane's being so open-minded and such a good learner that made it possible for his teacher to change him in the first place.

Here is another way to see roughly the same point. Suppose Zane has a classmate, Maria, who is almost exactly like Zane, but who is just a trifle less open-minded, and not quite as good a student, so that the college professor is not able to change Maria in any relevant way. Suppose that Maria, like Zane, ends up rushing into a burning school building and saving hundreds of children. In Maria's case, however, it will still be less than physically necessary that she will cause herself to run into the building to save the children. Thus, TAC4 yields a strange result concerning Zane and Maria. Zane, on this view, is not morally responsible for his heroic action (since his action is not morally free, according to the view), while Maria is morally responsible for her action (since her action is morally free, according to the view). So on this view, Maria's action is morally right, while Zane's is just morally neutral. This consequence seems clearly wrong. After all, if there is a moral difference between the two characters, it should be Zane and not Maria who enjoys a higher moral status – for Zane is in general more likely than Maria to cause himself to do the right thing, and Zane is also both a better student and more open-minded.

These considerations suggests a general argument against incompatibilism. Suppose it is true. Then presumably there are some factors other than a mere absence of physical necessity that are relevant to moral freedom. (For example, the proponent of the Theory of Agent Causation will say that the other main factor that is relevant to moral freedom is whether the action in question is caused by its agent.) And presumably there is some probabilistic cutoff such that the presence of those relevant factors plus a degree of "physical probability" below that cutoff results in a case involving moral freedom, while the presence of the relevant factors plus a degree of physical probability at or above the cutoff results in a case not involving moral freedom. (Perhaps the cutoff will be 1, or .7, or .5, or some other degree of physical probability.) But then consider two cases, alike in all respects (including the presence of the other factors that are relevant to moral freedom) except for the degree of physical probability involved. One is below the cutoff, but is arbitrarily close to being at it, while the other is right at the cutoff. One

case will involve moral freedom, according to the version of incompatibilism in question, and the other won't. But they will be arbitrarily close to being exactly alike. For example, if the cutoff is physical necessity (i.e., a degree of physical probability equal to 1), then in one case it will be physically necessary that the agent performs the relevant action, while in the other case it will not be physically necessary that she performs the relevant action, but it will be probable to degree .9999... , for some arbitrarily high number of places. And it is hard to see how the one case could involve moral freedom while the other doesn't, given that they are so alike.

An incompatibilist might respond to this continuum argument by denying that there is any sharp cutoff between morally free actions and actions that are not morally free. The incompatibilist might say instead that moral freedom comes in degrees, and that the difference between the two cases described in the continuum argument is that one involves no moral freedom at all, while the other involves a very small degree of moral freedom. But the thesis that there are degrees of moral freedom is a controversial one. So if it turns out that this is the best incompatibilist response to the argument, then the argument would at least show that incompatibilists ought to be committed to something controversial, namely, degrees of moral freedom.

Moreover, the degrees of freedom response would raise difficult questions about what makes for the highest degree of moral freedom. Is it that an agent is morally free to the highest degree when the relevant degree of physical probability is exactly .5? That seems arbitrary. Or is it that an agent is morally free to the highest degree when the relevant degree of physical probability approaches 0? But then to be morally free to the highest degree, you would have to do something that practically violates the laws of nature; and that is surely asking way too much.

Another problem with the degrees of moral freedom response to the continuum argument against incompatibilism is that it seems to entail that there are also degrees of moral responsibility. An agent whose action is morally free to degree .5, for example, would presumably be morally responsible for her action to degree .5, while another agent whose action is morally free to degree .1 would presumably be morally responsible for his action only to degree .1. Surely this way of talking about moral responsibility would be very odd.

A second possible incompatibilist response to the continuum argument would be to say that it is impossible for two cases to be alike with respect to

the presence of the other factors that are relevant to moral responsibility, yet different with respect to the degree of physical probability. But this response seems both *ad hoc* and pretty implausible. In any event, the incompatibilist who makes this response to the continuum argument would again be committed to something extra, and controversial.

6 A Compatibilist Version of the Theory of Agent Causation

Consideration of the above examples, together with the continuum argument against incompatibilism, suggests that the most plausible version of the Theory of Agent Causation will be one that does not require any other condition for an action's being morally free apart from the condition that the action is caused by its agent. In other words, we should simply get rid of the second conditions in all of the previous versions of the Theory of Agent Causation, and not replace them with anything. Our new version of the Theory of Agent Causation will then look like this.

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

COMTAC incorporates the fundamental insight behind the traditional versions of the Theory of Agent Causation – that an action is morally free only if it is caused by its agent. But since COMTAC lays down no other conditions for an action's being morally free, it manages to avoid the problems raised above for the previous versions of the Theory of Agent Causation. Thus, COMTAC gets the right results in Cases 2, 3, 4, and 5. The actions of Tim, Imran, Zoe, Mookie, and Zane are all morally free actions, according to COMTAC, and I think that these results accord with our intuitions about those cases.

A second advantage of COMTAC over its predecessors should be immediately apparent: COMTAC is remarkably neat and simple. On this view, there is just one condition to worry about when it comes to determining whether an action is morally free. If the action is caused by its agent, then it is morally free. Otherwise it isn't. Nothing could be simpler.

7 Some Objections to COMTAC

Near the beginning of this paper, I appealed to two intuitions about freedom and responsibility in order to motivate TAC1 and TAC2, the traditional versions of the Theory of Agent Causation. One of those intuitions was that

you can't be responsible for a spontaneous action, since you would have nothing to do with the causing of such an action. It should be clear that this intuition supports the main idea behind COMTAC. For the intuition is that you can't be morally responsible for an action that has no cause, and COMTAC – like every other version of the Theory of Agent Causation – implies that you can't be morally responsible for an action unless you are the cause of that action. But the other intuition appealed to in motivating the traditional versions of the Theory of Agent Causation involved the idea that you cannot be responsible for an action if it has causes outside of you, because in that case your action is, in a certain sense, beyond your control and, hence, not up to you. This intuition seems to go against any wholly compatibilistic version of the Theory of Agent Causation, such as COMTAC. Thus, it might be objected, one of the main intuitions that motivates the Theory of Agent Causation in the first place is an incompatibilist intuition that undermines COMTAC.

In response to this objection, I would say that what is really behind the idea that you cannot be responsible for an action if it has causes outside of you is not in fact a wholly incompatibilist intuition. Let me explain.

The reason it seems true that you cannot be morally responsible for an action that is caused by events outside of you, I would suggest, is that we tend to forget about the possibility of double causation. We forget that an action caused by some event outside of you could also be caused by you. So when we hear about an action that is caused by some event outside of you, we automatically think of an action that, in addition to having a cause outside of you, is not caused by you. Such an action would indeed be beyond your control, and it is hard to see how you could be morally responsible for it. But this intuition – that you could not be morally responsible for an action that is not caused by you – is of course consistent with COMTAC. In fact, it is the main idea behind COMTAC.

If you are still inclined to think that the intuition in question is an incompatibilist one, consider Case 3 again. That was the case in which Yasmine's asking Imran to pass the salt caused Imran's passing of the salt, but Imran himself, exercising agent causation, also caused his passing of the salt. It seems to me that once it is understood that Imran himself caused his own action, it is really quite plausible to think that Imran's action was morally free, even though it had a cause outside of its agent. Of course, things would have been different if Imran's action had been caused by Yasmine's request without also being caused by Imran himself. But that's not the way it went, and the

way it went – with Imran's action having at least two causes, one of which was Imran himself – seems consistent with the action's being morally free.

The upshot, I would claim, is that the Theory of Agent Causation is not in fact based on a genuinely incompatibilistic intuition. Rather, it is based on an intuition – that you cannot be morally responsible for an action that is caused by something outside of you, but not caused by you – that might appear at first glance to be incompatibilistic, but that, upon examination, turns out to be a compatibilistic intuition that supports the main idea behind COMTAC.

Another objection that might be raised against COMTAC involves cases like the following.

> **Case 7:** Roy is held up at gun point. He reaches into his pocket and reluctantly hands over his wallet. This action is caused by himself, through the exercising of the power of agent causation.

According to COMTAC, Roy's action in this case will be morally free, since it is caused by its agent. (According to what I have said above, it is plausible to think that Roy's action will also be caused by the gunman's actions. But according to COMTAC this is irrelevant to the question of whether the action is morally free.) But many people, including many compatibilists, have thought that actions such as this one are not in fact morally free. The idea is that we could hardly blame Roy for handing over his wallet, or say that what he did was morally wrong, since he was faced with the threat of being shot if he did not do what the gunman said. A.J. Ayer is an example of a compatibilist who takes this line. Ayer says,

> If I am constrained, I do not act freely. But in what circumstances can I legitimately be said to be constrained? An obvious instance is the case in which I am compelled by another person to do what he wants. In a case of this sort the compulsion need not be such as to deprive one of the power of choice. It is not required that the other person should have hypnotized me, or that he should make it physically impossible for me to go against his will. It is enough that he should induce me to do what he wants by making it clear to me that, if I do not, he will bring about some situation that I regard as even more undesirable than the consequences of the action that he wishes me to do. Thus, if a man points a pistol at my head I may still choose to disobey him: but this does not prevent its

> being true that if I do fall in with his wishes he can legitimately be said to have compelled me. And if the circumstances are such that no reasonable person would be expected to choose the other alternative, then the action I am made to do is not one for which I am held to be morally responsible.[12]

Thus it might be objected that COMTAC gets the wrong result in Case 7, and in similar cases involving coercion.

In reply to this objection, I would say that Roy's action is in fact morally free. After all, it would be perfectly natural to say that Roy did the right thing in this case. If he had refused to hand over his wallet, and had gotten shot for his stubbornness, we could plausibly say that he had done the wrong thing.[13] It's true that he was threatened with a bad consequence for performing an alternative action, so that he was not, at the time of his action, free from coercion. But freedom from coercion is a different matter from moral freedom.[14] If it's true that Roy did the right thing in this case, then he is morally responsible for his action. And if he is morally responsible for his action, then it follows that his action is morally free. So I think that COMTAC gets exactly the right result in this case.

In fact, I think it has been a big mistake on the part of philosophers like Ayer to deny that such actions are morally free, for the reason given in the above paragraph. But I have a diagnosis of what led these philosophers to make this mistake. My guess is that philosophers who have suggested that coerced actions are not morally free have done so because they were reluctant either to blame agents for performing actions under coercion or to say that

---

12 Ayer, "Freedom and Necessity," pp. 278-79.

13 Assuming, that is, that we fill out the example in a natural way: Roy is a good person, with a family that depends on him; there is no way for him to refuse to hand over his wallet and avoid being shot; the loss of the wallet will constitute only a minor inconvenience for Roy and his family; etc.

14 One important difference is that it seems clear that freedom from coercion comes in degrees, whereas it is at best a controversial matter whether moral freedom comes in degrees.

For a discussion of the difference between freedom from coercion and moral freedom, as well as some related issues, see Murray and Dudrick, "Are Coerced Acts Free?"

actions performed under coercion are morally wrong. It is understandable that these philosophers should have been reluctant in this way. But if this is what led the relevant philosophers to suggest that actions performed under coercion are not morally free, then those philosophers must have overlooked the possibility that an agent who performs an action under coercion might not be blameworthy, and the action in question might not be morally wrong, for the simple reason that the action is morally right.[15] And, as I have suggested, this is precisely how it is in the case of Roy. The upshot, I think, is that any plausible account of moral freedom will have to allow that coerced actions like Roy's can in fact be morally free.

Here is a third objection that might be made against COMTAC. It might be thought that COMTAC gets the wrong result in a case like the following.

> **Case 8:** Kid Patriot is an extremely earnest and patriotic secret service agent, who wants only to do a fine job of protecting the president. One night, Martians kidnap Kid and mess with his brain in a way that makes it physically necessary that Kid will cause himself to shoot the president. Then they return him to Earth, and he causes himself to shoot the president.

It is natural to think that Kid is not morally responsible for his action of shooting the president in Case 8 – the Martians are instead – and that his action is therefore not morally free. But according to COMTAC, Kid's action in Case 8 is morally free. So it might be thought that Case 8 refutes COMTAC.

My response to this objection is to accept the relevant consequences of COMTAC. Kid's action of shooting the president is morally free, and (assuming that there are no other mitigating circumstances) Kid is morally responsible for his action. This might be a case of biting the bullet, but I don't think it's a case of swallowing the cannon. Here's why.

Consider the case of a serial killer who performs his crimes as a result of being abused as a child. Suppose even that events in the killer's childhood that were out of his control forced him to be the way he is now, so that he now causes himself to kill people. It seems to me that even if this is true, it is still true that the serial killer is a bad person, and that what he does is morally wrong. We might not blame the serial killer as much, if we know that an abuser made him the way he is. But that doesn't mean that what he does is

---

15 *Cf*. J.L. Austin's remarks about "justifications" in his "A Plea for Excuses."

not morally wrong. And of course if his actions are morally wrong, then they are morally free.

I think it is the same with Kid Patriot. The Martians made him a bad person. It was wrong of them to do that, and they shouldn't have done it. But they did. Here it is extremely important to notice that in order to make sure that Kid would cause his own action, the Martians had to change Kid, the same way the abuser had to change the serial killer before the latter would cause himself to kill people. It was not enough merely to cause an event in Kid's brain that would cause the shooting of the president (perhaps by putting some foreign object in Kid's head that would zap a certain neuron at just the right time). To force Kid *to cause himself* to shoot the president, the Martians literally had to change his very nature, so that *he* would do the relevant causing. Presumably they were able to do this by changing things around in his brain. But the point is that they had to change *him* in a way that is not analogous to forcing him to lift his arm by doing it for him.

Thus, even if Kid used to be a good person, now that the Martians are through with him this is no longer true. Kid is now a bad guy who has done a bad thing. It's true that he became a bad guy through no fault of his own, but the above problems with TAC4, together with the continuum argument, suggest that we should not make it a requirement, for an action to be morally free, that its agent be responsible for his own character if that character makes it physically necessary that he cause himself to perform that action. Here is another case that seems to suggest the same thing.

> **Case 9:** A brand new person, Leila, pops into existence one morning. She has a fine moral character that makes it physically necessary that she will do a certain good deed in a certain situation. The situation arises and, sure enough, Leila performs the good deed.

It seems to me that Leila deserves praise for her action in this case, and also that her action is morally right. So it seems to me that Leila's action should be considered morally free. But of course Leila is not in any way responsible for her character, and that character makes it physically necessary that she perform the action in question.[16]

---

[16] If I am right about Leila's action in Case 9 being morally free, then this case is another counterexample to TAC4.

Returning to the case of Kid Patriot, notice that even if we say that Kid is morally responsible for shooting the president, we can still absolve Kid of any blame for becoming a bad person, and this perhaps accommodates at least part of the intuition that Kid does not deserve blame in Case 8. The fact that Kid deserves no blame for becoming a bad person might also affect our thinking on what punishment Kid deserves for his crime. We might be more inclined to try to rehabilitate him than we would have been if he had turned himself into a bad person, and we might think that he deserves a chance at rehabilitation more than he would have if he had been responsible for becoming a bad person. But the bottom line is that, in light of all the problems for compatibilist versions of the Theory of Agent Causation, it seems best for the proponent of the Theory of Agent Causation to endorse COMTAC and bite the bullet in cases like that of Kid Patriot.[17]

8 Why Has the Theory of Agent Causation Always Been Thought of as an Incompatibilist Theory?

Unless there is some serious flaw in the above arguments, it has been relatively easy to show that the most plausible version of the Theory of Agent Causation is a wholly compatibilist one. Given this, it may strike the reader as somewhat surprising – if not downright odd – that the theory has, until now, always been thought of as an incompatibilist theory. I agree that this is indeed an odd phenomenon, but I suspect that there is a perfectly good explanation for it.

The explanation I have in mind has to do with contingent, biographical facts about the various people who have endorsed the Theory of Agent Causation. My guess is that something like the following happened. Before they became proponents of the Theory of Agent Causation, these people first became persuaded, by familiar kinds of arguments, that incompatibilism is true. Then they worried about the problem of reconciling moral responsibility

---

[17] It might be thought that there is room to wiggle on the matter of whether Kid's action satisfies the other conditions necessary for his being morally responsible. (See footnote 3 above.) But I don't think this approach will work, for it seems clear that if there is any reason at all why Kid would not be morally responsible for his action, it would have to be that, as a result of what the Martians did to him, his action is not morally free.

with indeterminism. If we want to say that an agent is morally responsible for an action, and we are worried about incompatibilistic considerations, it doesn't seem to help much just to say that the action is uncaused (or not physically necessary, or whatever). How can the agent be responsible for her action if nothing – not she or anything else – causes that action? Finally, the relevant people decided that in order for an agent to be morally responsible for an action, the agent would have to be the cause of her own action.

In other words, I suspect that the people in question were incompatibilists first, and then their incompatibilism drove them to the Theory of Agent Causation. Once they endorsed the Theory of Agent Causation, they neglected to notice that the most plausible version of it is a compatibilist one.[18]

9 Conclusion

Cases 2, 3, and 4 above show that the traditional versions of the Theory of Agent Causation are untenable. They also show that the Theory of Agent Causation should not entail that the Free Will Thesis is incompatible with either the Principle of Universal Causation or Event Determinism. Cases 5 and

---

[18] Something like the first part of this story is what happened to me. I was brought up (by my graduate school teachers, and by reading people like David Lewis (see his "The Paradoxes of Time Travel" and "Are We Free to Break the Laws?")) to be a compatibilist. But then I read van Inwagen's *An Essay on Free Will*, and changed my mind. I became an incompatibilist. But I still didn't see how agents could be free, unless they were the causes of their own actions. So I became drawn to the Theory of Agent Causation. Finally, I realized that if I was going to subscribe to a version of the Theory of Agent Causation, then I should subscribe to a compatibilist version of the theory.

Speaking of van Inwagen's book, the reader might wonder how the proponent of COMTAC ought to respond to some of the well-known arguments for incompatibilism, including van Inwagen's powerful arguments. My own view is that the proponent of COMTAC ought to avail him- or herself of the standard compatibilist responses to those arguments. In fact, I am inclined to think that in many cases, the standard compatibilist responses to the incompatibilist arguments are much more plausible on the assumption that agents can cause events than they are without that assumption. But in any case, this topic – of the standard arguments for incompatibilism, and how the proponent of COMTAC ought to respond to those arguments – is beyond the scope of the present paper.

6 show that versions of the Theory of Agent Causation that entail that the Free Will Thesis is incompatible with Future Determinism, the strongest version of determinism, are also untenable. Moreover, consideration of the continuum argument discussed in Section 5 above adds independent support to this claim. And if what I say about Cases 7, 8, and 9 is correct, COMTAC, the completely compatibilistic version of the Theory of Agent Causation presented here, successfully avoids the problems facing the traditional versions of the Theory of Agent Causation while allowing its proponents to give plausible replies to objections against it. I conclude that COMTAC is the best available version of the Theory of Agent Causation, and that, contrary to what people have thought in the past, the most plausible version of the Theory of Agent Causation is a compatibilist theory rather than an incompatibilist one.[19]

---

[19] Earlier versions of this paper were presented at Western Washington University, the Pacific Division of the American Philosophical Association (in March, 1997), Notre Dame University, and West Virginia University. I'm grateful to the people present on those occasions – especially Keith Butler, Thomas Downing, Neil Feit, Frances Howard-Snyder, Hud Hudson, Phillip Montague, Loretta Torrago, Peter van Inwagen, and Dean Zimmerman – for helpful criticism. I'm also grateful to Mark Aronszajn, Theodore Drange, Fred Feldman, Richard Feldman, Ishtiyaque Haji, Mark Heller, Frances Howard-Snyder, Sharon Ryan, Theodore Sider, Donald Turner, Kadri Vihvelin, and Linda Wetzel for helpful comments on earlier versions of the paper.

References

Aristotle, *Eudemian Ethics*.

Aristotle, *Nichomachean Ethics*.

Aristotle, *Physics*.

Austin, J.L., "A Plea for Excuses," in Austin, J.L., *Philosophical Papers* (Oxford: Oxford University Press, 1961).

Ayer, A.J., "Freedom and Necessity," in Ayer, A.J., *Philosophical Essays* (New York: Macmillan and Co., 1954).

Campbell, C.A., "Is 'Free Will' a Pseudo-Problem?" *Mind* **60** (1951).

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)).

Chisholm, Roderick, *Person and Object* (La Salle, IL: Open Court Publishing Co., 1976).

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).

Frankfurt, Harry, "The Principle of Alternate Possibilities," *The Journal of Philosophy* **66** (1969).

Lewis, David, "Are We Free to Break the Laws?" in Lewis, David, *Philosophical Papers, Volume II* (Oxford: Oxford University Press, 1986).

Lewis, David, "Causation," in Lewis, David, *Philosophical Papers, Volume II* (Oxford: Oxford University Press, 1986).

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

Murray, Michael J., and Dudrick, David F., "Are Coerced Acts Free?" *American Philosophical Quarterly* **32** (1995).

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

Reid, Thomas, *Essays on the Active Powers of Man*.

Suarez, Francisco, *Disputationes Metaphysicae*.

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

Taylor, Richard, "Determinism and the Theory of Agency," in Hook, Sidney (ed.), *Determinism and Freedom in the Age of Modern Science* (New York: Collier Books, 1961).

Taylor, Richard, *Metaphysics* (Englewood Cliffs, NJ: Prentice-Hall, 1992).

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