

Freedom Cancellation and Determinism

Alexander R. Pruss
January 31, 2004

e-mail: *ap85@georgetown.edu*
Department of Philosophy
Georgetown University
Washington, DC 20057

1. Introduction

Compatibilists believe that our intuitions that free will and determinism are incompatible can be explained by the fact that we have an unjustifiable tendency assimilate determinism to situations that are genuinely freedom canceling. To do justice to these intuitions, the compatibilist generally admits several kinds of cases of freedom cancellation but insists that the case of determinism is significantly different from these cases. A very common example of a freedom canceling condition is when one person's actions are solely the result of another's causal control. Most compatibilists think that in such a case, the person performing the actions is not responsible for them, but the responsibility rests exclusively in the person exercising the control.

I will argue that, compatibilist orthodoxy notwithstanding, once one admits that causal control of one person's actions by another person is freedom canceling, then one should also admit that determinism would be freedom canceling. If the argument is sound, then either one should abandon the general principle that control of one person's actions by another is always freedom canceling, or else one should become an incompatibilist. I take the latter to be the more attractive option.

The specific freedom cancellation principle I will be interested in is the Total Personal Control Principle, defended in somewhat different form by Richard M. Gale (??ref):

(TPCP) If there is a person x who intentionally and deterministically brought about *all* of the actions of y , then none of y 's actions are free.

That the control is over *all* of y 's actions is significant. Richard Gale in conversation has argued that if he knew I was a nice person who would do him a small favor when asked for it, his asking me for that favor would not be freedom canceling, even though the request would have intentionally brought about

my action. It is only if the control is over *all* of a person's actions that it is uncontroversial that we have a case of freedom cancellation.

The notion of *deterministically* bringing about an event is also important. I say that x deterministically brings about an event F providing that x directly and intentionally causes an event E which when conjoined with the circumstances C occurrent at the time of x 's action is such as to nomically necessitate F 's occurrence. This, furthermore, counts as intentional provided that x intends that F should result in this way from E and C . Without the restriction to deterministic causal control, TPCP would be less plausible, even to a libertarian. Imagine, after all, a very good libertarian-free agent who is still predictable—it is very likely that she will choose the right thing. Suppose, too, that the agent is quite short-lived, so that she only engages in three actions in her life. Suppose in each case Fred places the agent in circumstances in which the right decision will be the one he want her to make, and that she in fact makes the right decision. Then Fred has intentionally brought about all of her actions, but she could have done otherwise, and there is at least some plausibility to the idea that she acted freely.

I take TPCP to be plausible common ground between a compatibilist and incompatibilist. Some people may think I am too cautious in my formulation of TPCP, and they might be right. But as it stands, TPCP is hard to deny. I will argue, however, that if one is set on accepting TPCP, one should abandon compatibilism.

I am going to make my main argument in two different ways, by choosing two different kinds of premises to conjoin with the TPCP. I only need one of the two ways to succeed. Then I will discuss ways in which a compatibilist might weaken the TPCP.

Throughout I will make the simplifying assumption that a compatibilist believes that “human-like” persons can be free. My standard for “human-likeness” is going to be very loose, as the foregoing will show, but will include such characteristics as having a finite past that starts significantly after the beginning of the universe. Someone who held that if determinism held an *eternal* person could be free would not count as a compatibilist on my account.

2. First argument: Supervenience of freedom on recent history

The “recent history” of a human-like person x in a world w will, by stipulation, be that period of the history of w that includes all of the life of x as well as some time before that, but not going back to the

beginning of the universe if the universe has a beginning. It will not matter exactly how much “some time before that” is. If human-like persons are posited to be mortal, we can define it as an amount of time equal to x ’s life-span.

This lets us formulate the following Recent History Supervenience Principle:

(RHSP) If x is a human-like person in a world w and y^* is a person in a world w^* who lives at the same times as y does, and the worlds w and w^* match over the recent histories of y and y^* respectively, with y and y^* corresponding, then y freely acts at t in w if and only if y^* freely acts at t in w^* .

This principle depends on a notion of the *matching* of worlds over a set T of times that probably cannot be defined with complete precision. The idea is that the two worlds’ individuals and their interrelationships as restricted solely to what is happening during T are indiscernible. Somewhat more precisely, w and w^* match over T providing there is a one-to-one map f that maps an individual substance or event of w existing or occurring at some time in T to an individual of w^* existing or occurring at some time in T , such that:

- (a) every individual of w^* existing or occurring at some time in T is of the form $f(x)$ for some x existing or occurring in w at some time in T ;
- (b) if R is a T -pure n -ary relation (i.e., property if $n=1$) that is purely qualitative except possibly for rigid references to times and places, then $Rx_1\dots x_n$ holds at w if and only if $Rf(x_1)\dots f(x_n)$ holds at w^* .

Moreover, we say that two individuals y and y^* *correspond* providing that $y^*=f(y)$ for some such map f .

The above definition of “matching” still depends on two primitive notions. One is that of a relation that is purely qualitative except possibly for rigid references to times and places. Roughly, this is a relation that can be stated without making use of demonstratives or proper names for anything other than possibly times and places, but since for all we know there might be purely qualitative relations that cannot be stated at all in a discursive language, this is not a precise definition. The second notion is that of a T -pure relation. Roughly, inspired by Richard Gale’s notion of a temporally pure proposition(??ref), we might say that a T -pure relation R is one such that if $x_1\dots x_n$ are names of individuals that exist at some time in T , then the proposition that $Rx_1\dots x_n$ does not entail the existence of any times outside of T and is compatible with any number of repeated tokenings of “ $Rx_1\dots x_n$ ” being true. This, again, is not a precise definition^[1], though

it may work for logically simple relations. The basic idea is that whether a T -pure relation holds between a bunch of individuals does not depend on what happens to these individuals outside of T .

Despite the lack of precision, I think the intuitive idea is clear. Moreover, one needs something like this in order to formulate the notion of determinism, since we would say that laws L are deterministic providing that if worlds w and w^* have laws L , and w and w^* match over the set of all times up to and including t , then w and w^* match over the set of all times.

RHSP is highly plausible. Whether Curley *freely* took a bribe at t should not depend on what happened a length of time before Curley's conception. Nor should it depend on Curley's numerical identity: anybody indiscernible from Curley who acted indiscernible is free if and only if Curley is.

But the conjunction of RHSP, TPCP and compatibilism is incoherent. To see this, suppose that compatibilism holds. Let y be a human-like person who acts freely at t in a deterministic world w . Let T be the recent history of y . Let w^* be a world that matches w over T , but where there is a very smart and very powerful alien who shortly before T intentionally engineered the physical conditions in our galaxy precisely with a view to producing a person, y^* , who is just like our world's y , and who acts precisely the way y acts in our world. By the RHSP, y^* acts freely at t if y acts freely at t . But the TPCP would imply that y^* never acts freely. Hence, the conjunction of compatibilism and RHSP is incompatible with TPCP. Given the deep plausibility of RHSP, someone who accepts TPCP sufficiently strongly should reject compatibilism and someone who has a sufficiently strong commitment to compatibilism should reject TPCP.

3. Second argument: Unintentional action, zombies and the like

The second argument proceeds by citing a series of intuitions which imply that TPCP is unnaturally restricted, and thus move us from TPCP to incompatibilism.

Step 1: Unintentional action. If my actions are unfree precisely as a result of the influence your actions have on me, then intuitively what is relevant to the evaluation of my responsibility is not what *you* were thinking, but what effect your actions have had on me and how they impacted *my* thinking.

Suppose, for instance, that you point a gun to my head and say: "If you do not go and rob the store, I will kill you." Whether I was responsible for robbing the store does not depend on whether you intended this to be a joke (perhaps you thought you had told me ahead of time that you would make this joke, but in

fact you had not) or whether you were serious, except insofar as you might have used a different tone of voice or had a somewhat different face in the two cases. But if the difference between the case of a joke and the case of a serious threat were not reflected in your outward actions at all, then the question whether you were joking or serious is irrelevant to the evaluation of my responsibility, though highly relevant to the evaluation of yours.

If we accept this intuition, then we should drop the intentionality from TPCP, and retain only:

(TPCP-2) If there is a person x who deterministically brought about *all* of the actions of y , then none of y 's actions are free.

For suppose that we accept TPCP and the intuition that your intentions are irrelevant to the evaluation of my responsibility except insofar as your intentions are outwardly manifested. Suppose, further, that you deterministically brought about all of my actions, but did not do so intentionally. Then it was logically possible for you to have done this intentionally but with the same outward movements. For it was logically possible for you to know (e.g., because an alien genius communicated this to you telepathically) that such-and-such a sequence of outward movements would produce such-and-such actions in me, and it was thus logically possible for you to desire to produce these actions by these movements. But if the evaluation of my responsibility does not depend on your intentions but only on your outward actions, then I am either free in both or neither case. Since the TPCP entails that in the case of intentional control I am not free, it follows that I am free in neither case, and TPCP-2 follows.

Objection. Soldiers in a war are not responsible for their actions precisely when they follow the orders of their superiors. But whether A was ordered depends on the intention of the superior—"Shoot the spy" is not an order if muttered in the commander's sleep. Hence, my responsibility may be affected by your intentions.

Response. First, I deny the general claim that soldiers are not responsible when they follow the orders of their superiors. The *Uniform Code of Military Justice* (??ref) expressly prohibits soldiers from following immoral orders. Second, if the claim were true, then I would bite the bullet: it should make no difference to whether a soldier was responsible for an action whether the commander *actually* commanded it or merely *seemed* to command it, as long as the seeming command looked exactly like an actual command.

Step 2: Zombies. We have so far the claim that if you acted in such a way as deterministically brought about all of my actions, then I am not free regardless of what you might have been thinking. But if so, then it seems that even if you were not thinking at all, and influenced me by “actions” done unconsciously in your sleep, I should still be unfree. For there does not seem to be any reason why it should make a difference to *my* responsibility whether you brought about an effect unwittingly and unintentionally, or whether the effect came from something entirely unconscious in you and relevantly similar to a twitch. Our previous intuition should extend to this: Not only the exact content of what is going on in your head should not affect my responsibility, but neither should my responsibility be affected by whether there was any content to what was going on in your head.

Consequently, if you deterministically brought about all of my actions while you were sleep-walking, I would still be unfree. Moreover, there should be no difference for my responsibility between the case of my being causally influenced by you while you were sleep-walking and the case of my being causally influenced by a zombie who looked like you. Thus, even if a zombie deterministically brought about all of my actions, still I would be unfree, as long as he did it in a way outwardly indistinguishable from the way a person might have done so.

Since anything done by a zombie could be consciously done by a person with *some* purpose in mind (the purpose might just be to look like a zombie!), we can extend TPCP-2 to include that which is done by zombies. Note that *if* a soldier is not responsible for actions done in response to orders, neither would the soldier be responsible for actions done in response to a zombie who looked and acted just like the commander.

Step 3: All other causes of causal determination. But since a zombie is actually just an object and not a person, there should be no difference for my responsibility based on whether my actions were caused deterministically by a zombie or by other objects. If this intuition holds, then from the last extended version TPCP, we can conclude:

(TPCP-3) If all of y 's actions were deterministically caused by things outside of y , then none of y 's actions are free.

There is another way to this conclusion from the zombie case. There is no one way for a zombie to look and be. A zombie is something that looks and behaves just like a person might but is not a person.

Now, there are few logical restrictions on what a person looks and behaves like outwardly. A person could look and behave outwardly like a swamp, interacting in subtle causal ways through its waves with another person who looks outwardly like a tree. This is not a dualist claim. By *outwardly*, I mean “outwardly” in the crude physical sense. The swampish person might have a physical brain somewhere in the middle of the swamp, but I am assuming that this fact is not outwardly discernible from my point of view. A zombie-human by my lax definition need not have a brain at all, as long as its causal effects on its environment are the same as that of a human with a brain.

But if there are few logical restrictions on a person’s outward appearance and behavior, then just about anything is, logically speaking, a zombie. An ordinary swamp is a zombie, i.e., something that looks and acts outwardly the way a logically possible swampish person would. Hence the restriction in the extended version of the TPCP to the case of actions caused deterministically by a zombie is not much of a restriction at all. Perhaps a person could not outwardly look just like an electron—perhaps more physical complexity is needed to constitute a person (though a dualist might disagree). But as soon as one has sufficiently many particles, and hence has enough room to “hide” a brain-like organ, one has something that outwardly looks and behaves just like some logically possible person does. And to restrict the TPCP to cases where the cause has sufficient complexity that it could “hide” a brain-like organ would be unnatural: surely if the entity doesn’t *have* a brain, it doesn’t matter vis-à-vis a general freedom cancellation principle whether it has room to hide one.

If this is right, then once we have accepted the extension to actions being caused by a zombie, it seems we should accept TPCP-3. The one objection that one might make is that TPCP-3 includes the case where *y*’s actions are caused by a number of scattered causes, while a zombie is perhaps not a scattered individual. However, any earthly case of causation by a number of scattered causing individuals can be reconceived of as a case of causation by a larger non-scattered individual that includes the causally irrelevant air molecules between the scattered causes. Moreover, there does not seem any obvious reason why a zombie couldn’t be a scattered individual, since a person could logically be such, e.g., with the various parts in the person’s body possibly communicating in subtle ways through gravitational radiation. Finally, there does not seem to be any reason to think that one is any the more free when one’s actions are caused by a committee of zombies than if they are caused by one zombie.

Step 4: Determinism simpliciter. I have so far argued that acceptance of the TPCP, together with plausible intuitions, should push us to an acceptance of TPCP-3, which essentially says that freedom is incompatible with causal determinism, at least for human-like persons since a human-like person in a causally deterministic universe has all of her actions be deterministically caused by events prior to her existence, as a human-like person has a beginning in time not coinciding with the beginning of time. This is close to incompatibilism, except that it leaves open the possibility that a human-like person might be free in a deterministic universe where the determinism is non-causal: where it is not the case that later states are deterministically *caused* by earlier ones.

That said, it would be a strange kind of a view on which non-causal determinism is compatible with free will but causal determinism is not. It would seem that a non-causally deterministic universe would be one where there would not be causation at all between events or non-divine substances.^[2] For causation would seem to be explanatorily *de trop* at such a world—non-causal deterministic explanations would seem sufficient. And this would seem to threaten free will even more.

Thus, it seems that we are inexorably pushed from accepting TPCP to accepting incompatibilism.

4. Conclusions

The basic argument here was based on the intuition that whether a person is responsible for her actions should not depend on what is happening significantly outside the person, whether in the head of another person or in the distant past. This is a kind of internalism about responsibility.

Those who reject such internalism because they think that responsibility is something that cannot be predicated of an individual in isolation from a community will reject my second argument, since that argument starts from the assumption that whether I am responsible does not logically depend on what is going on in the mind of another. However, such a view is still compatible with the first argument, as long as we strengthen the notion of a human-like person to be one that not only came to exist after the start of time but whose community came to exist after the start of time. For we can then extend RHSP to a communal version:

(RHSP-C) If x is a human-like person in a world w and y^* is a person in a world w^* who lives at the same times as y does, and the worlds w and w^* match over the recent histories of the communities

of y and y^* respectively, with y and y^* corresponding, then y freely acts at t in w if and only if y^* freely acts at t in w^* .

Of course someone might reject internalism to such an extent that no form of RHSP or of my intuition that freedom doesn't depend on what is happening in other people's heads will be acceptable. I have little to say to a person who opts for such an implausible view.

The reasonable option, I think, is to choose between TPCP and compatibilism. Were I a staunch compatibilism, I would reject TPCP in favor of a weaker principle that says that if you deterministically cause all of my actions *by doing A, B or C* (where the list would likely be much longer), then I am not free. I would then accept that if a zombie or a force of nature did *A, B or C*, I would not be free either. However, I would maintain, that as a matter of fact the genetical-environmental determination that would obtain in a world like ours if determinism held would not be of the form of *A, B or C*. Here, the ways of acting I label "*A, B and C*", would include such things as brainwashing and paralyzing threats. Unfortunately, any such view would appear *ad hoc*, because it seems we could never be sure in a principled way that "*A, B and C*" list all the freedom-canceling ways that one person could act on another. Thus, TPCP seems theoretically preferable.

The only I can see of getting around the *ad-hocness* objection is to abandon the TPCP altogether, and simply hope that all cases of freedom cancellation due to the action of another person are to be analyzed by assimilating them to some *other* sort of freedom cancellation that a compatibilist would accept, such as freedom cancellation due to physical constraint, paralyzing fear, ignorance or inability to deliberate. Doing so would, however, make compatibilism less plausible. For there does seem to something deeply plausible about the idea that if in some clever way you made yourself the author of all of my actions in a causally deterministic way, then not only are you responsible for my actions, but I am not responsible for them.

^[1] Let F be the property of *being blue at some time outside of T or being green at some time inside T* . Intuitively, F should not be T -pure, but by this definition it is.

^[2] One might suppose such a world created by a timeless deity, and if so that creation might be the only instance of causation.