Philos Stud (2018) 175:1545–1556
https://doi.org/10.1007/s11098-018-1060-4

Replies to critics

Carolina Sartorio[1]

Published online: 21 February 2018

Abstract I respond to the critical comments by Randolph Clarke, Alfred Mele, and Derk Pereboom on my book *Causation* and *Free Will*. I discuss some features of the view that our freedom is exclusively based on actual causes, including the role played in it by absences of reasons, absence causation, modal facts, and finally some additional thoughts on how a compatibilist can respond to the manipulation argument for incompatibilism.

Keywords Free will · Causation · Compatibilism · Reasons-sensitivity · Absence causation · Manipulation

1 Reply to Clarke

Clarke first focuses on the concept of *acting on the basis of absences of reasons*, which plays a central role in my account of reasons-sensitivity. He draws a distinction between two conceptions of reasons, reasons as psychological states and reasons as facts, and examines how my view could be formulated in terms of each of these conceptions. This is helpful because it illustrates how the details of the view could be fleshed out depending on one's views on what reasons are. Clarke points out that, if one takes the view that reasons are facts, then acting for a reason arguably requires recognition of that fact; in contrast, acting on the basis of the absence of a reason doesn't require recognition of the relevant absence: all it requires is the *lack* of recognition of some fact (and the agent's conduct being caused, in a non-deviant way, by such a lack of recognition). Thus, on the view that reasons are facts, recognition (a psychological state of "mentally registering" some

✉ Carolina Sartorio
sartorio@email.arizona.edu

[1] University of Arizona, Tucson, USA

fact) doesn't play the same role in acting on the basis of absences of reasons that it does in acting for reasons.

This sounds right. But note that the Davidsonian view (the view that reasons are psychological states) also results in an asymmetry concerning the role played by psychological states. In this case, the asymmetry is that, whereas acting for reasons requires a certain causal role played by psychological states such as belief/desire pairs, acting on the basis of the absence of reasons doesn't: all it requires is a certain causal role played by the *absence* of psychological states. So either way there is an asymmetry between acting for reasons and acting on the basis of the absence of reasons.

Clarke then considers my claim that an agent's freedom or reasons-sensitivity is exclusively a function of actual causes. Is it really plausible to believe this? First, isn't a degree of sanity or contact with reality required? Arguably yes, but notice that this is consistent with the view because it can be incorporated as an additional requirement on the actual causes: the requirement that the reasons and absences of reasons that are the actual causes form understandable patterns of reasons-responsiveness. Although this is an additional requirement, it is consistent with the claim that freedom supervenes on actual causes (see *Causation and Free Will,* p. 138).

Clarke also suggests that we can act freely in some cases where we fail to be sensitive to reasons or absences of reasons, if the relevant facts are inaccessible to us (and not due to a deficiency on our part). In response, note that, in the case of actual reasons, it's common to understand our reasons-sensitivity in terms of the reasons that we *have*, which limits the relevant reasons to those that are accessible to us. As for absences of reasons, we can understand our sensitivity to them simply in terms of our sensitivity to our *lacking* certain reasons. If I decide not to carry an umbrella, this might be because I actually have reason to think that it's not raining, or because I'm less informed about the facts and I just lack a reason to think that it's raining. Either way, I can be sensitive to reasons, on the view that I'm proposing.

Clarke also discusses cases where an agent's negligence would have prevented her from recognizing and acting on absent reasons. In such cases, the absence of those reasons seems to play no role in what the agent actually does; still, Clarke suggests that, if the agent was *able* to recognize and act on the reasons, she can still act freely. In response, it seems to me that, in order for our agent to have the relevant ability, this *must* somehow be reflected in the actual causes. Imagine that I step on your foot, not noticing that it's there. Imagine that my negligence is so generalized that I would have failed to react even if you had been yelling at me not to advance, even if there had been a flashing sign in front of me pointing to the presence of your foot, etc. In this case it seems that I'm not able to respond to reasons of the relevant kind. So this suggests that, in the type of case that Clarke is imagining, where the agent *is* so able, this will have to be reflected in a range of absences of reasons, such as the absence of flashing signs, that she's *actually* responding to. Otherwise, the claim that she has the relevant ability seems unfounded.[1]

[1] Clarke (personal communication) notes that this still doesn't capture the idea that I'm sensitive to reasons because I'm sensitive to *the presence of your foot* (which I am because I'm able to recognize and respond to the presence of your foot, although I don't in the actual case). But it seems to me unclear that we really want to say this. Plus, it clashes with the lesson I draw from Frankfurt, since the presence of your foot doesn't play an actual causal role.

Clarke then focuses on my commitments about absence causation and overdetermination. On the one hand, I claim that, in certain cases, the fact that an outcome is overdetermined results in an agent's omission not causing the outcome. An example is the Accomplice case, where the agent who witnesses the robbery doesn't even try to call the police (but the robber's accomplice would have cut off the phone lines had he tried to call them): in this case the agent's omission doesn't cause the outcome of the man being robbed, which he couldn't have prevented. On the other hand, I claim that in other cases absences *can* cause outcomes that are overdetermined. Notably, in a Frankfurt case, the absences of certain reasons cause the agent's choice, which was overdetermined due to the presence of the neuroscientist. For example, if Frank is inclined to refrain from shooting victims who have kids, then the absence of reason R: Furt has kids, is among the causes of Frank's choice to shoot Furt, despite the fact that he would have ended up making the same choice if Furt had had kids (because the neuroscientist would have forced him to). But, what grounds the difference between these causal claims? Why is the absence of R a cause of Frank's choice to shoot Furt in the Frankfurt case, if the agent's omission is not a cause of the outcome in the Accomplice case? This is a fair question. As Clarke notes, I didn't provide an account of absence causation that yields these results, and I was partly relying on intuition to support these claims. But, is there anything else that could be said in this connection?

There is only so much I can do in the limited space I have here, but I'll try. (This is an interesting issue that deserves more consideration and that hasn't been sufficiently explored in the literature.) First, it's important to note that some extant accounts allow for the causation (or quasi-causation) of overdetermined outcomes in some special cases. Interestingly, Phil Dowe's view is one of them.[2] Dowe argues that absence-causation claims should be reinterpreted as claims about actual and counterfactual causal processes. If a child suddenly runs toward the road and is hit by a speeding car, and if the child's father could have warned the driver to stop but didn't, then, on Dowe's account, the father's failing to warn the driver quasi-caused the accident because, if he had warned the driver, this would have interrupted the causal process that actually led to the accident. Dowe argues that this is the case even if, had he warned the driver, another car would have still run over the child in a similar way (that is, even if the accident was overdetermined). For, even in that case, the father contributed to the process that *actually* led to the accident, by failing to warn the driver of the first car.

Interestingly, Dowe's account seems to result in a difference between Accomplice and the Frankfurt case. In the Frankfurt case, if Furt had had kids, Frank's considering this reason would have interrupted the process that actually led to his choice (his actual process of deliberation). Another process, initiated by the neuroscientist, would have then led to the same choice by Frank. But, just like in the

[2] I have in mind the account of quasi-causation offered in his book (Dowe 2000, chapter 6). In fact, David Lewis's view (Lewis 2004, the other account mentioned by Clarke) is also an example of this. While Lewis's account of direct causation requires strict counterfactual dependence, Lewis takes causation to be the ancestral of direct causation, which allows for causation of overdetermined outcomes in some cases.

accident case, this means that the absence of the reason made a contribution to the process that *actually* led to Frank's choice. So, on Dowe's view, the absence of the reason quasi-caused the choice. In contrast, the agent's trying to call the police in Accomplice wouldn't have interrupted the process that actually led to the man being robbed (the process initiated by the thugs). Thus, on Dowe's view, the agent's failure to try to call the police didn't quasi-cause the outcome.

Now, I'm not sure that I buy Dowe's account. Is it really plausible to believe that the powers of an absence hinge on whether there is a single process that cannot be stopped, as opposed to an actual process that can be stopped plus a "backup" process that would have led to the same outcome? This is far from clear. So I'd rather not rely on an account of this kind. But it's worth noting that at least one extant view about the powers of absences seems to have this result.

Another possibility is to say that what accounts for the causal difference between the cases is the nature of the effect itself: whether it's internal to the agent (a mental event like a choice) or external. It's plausible to think that causal relations can be more sensitive to external facts, such as facts that break the counterfactual dependence between two events, when we're dealing with external outcomes than when we're dealing with internal outcomes. Imagine a conscientious driver, Randy, who makes the choice to advance when the light turns green. When Randy makes that choice, he's responding to several features of the environment, including both positive and negative features: the light being green, the existence of a road up ahead, the *absence* of cars about to run the red light, the *absence* of fire-truck sirens, etc. Arguably, adding the neuroscientist in the background (who would have forced him to make the same choice to advance) doesn't change this: intuitively, Randy is still responding to all of these features, including the absences, when he makes the choice on his own. Now imagine that later on, when Randy gets to the intersection where he usually turns to go to work, he spaces out and misses the turn. He then takes a longer route and is late for work. Imagine, however, that the streets were closed off for construction, so he wouldn't have been able to take the shorter route anyway. Did his failure to turn at the intersection still contribute to his being late to work? It doesn't seem so.

So perhaps an absence of a reason can result in a choice when the choice is overdetermined because agents have certain dispositions—dispositions to respond to reasons and absences of reasons whose activation conditions don't depend on the choice not being overdetermined. Perhaps the causal powers of other absences are different in that they aren't grounded in such dispositions. At any rate, I agree that more could be said about this. For what it's worth, I take it to be an advantage of the view that I propose that it doesn't take a stand on what the grounds for causation are, including the grounds for absence causation or quasi-causation, since this results in a more neutral view that could be attractive to more people (I return to this point below).[3]

[3] Also, some causal claims that I make in the book are somewhat tentative. In particular, in chapter 2 I note that, if I'm wrong about the causal structure in Accomplice, the relevant claims about responsibility may have to be revised as a result (pp. 83–84). So, at the end of the day, I'm not completely wedded to every causal claim I make in the book. My most central commitment is to the idea that the responsibility facts are grounded in the causal facts, whatever these turn out to be.

Finally, Clarke raises some important questions about the grounds of freedom. First, if absences are nothing, then facts about agents responding to absences of reasons cannot consist in the obtaining of certain causal relations between absences of reasons and agents' choices (since some of the "relata" are missing). Then, what *do* they consist in? Also, assuming agents respond to reasons and absences of reasons in virtue of their having certain powers or dispositions, doesn't freedom remain something modal, in the end?

In response, let me note that, even on the view that absences of reasons are nothing, we can still think about the agent and her relation to the world, and ground her freedom on that basis. A free agent is someone who navigates the world in a certain kind of way, partly by attending to what is the case and partly by attending to what is not the case ("positive" and "negative" features about the world). As we have seen, there are different ways of understanding this claim from a metaphysical perspective. The simplest way assumes that absence causation is possible, and appeals only to actual causal relations. Other ways appeal to counterfactual causal relations as well as actual ones. But, even if the view ends up appealing to some counterfactual facts (or other modal notions like dispositions or powers of some kind), the counterfactual facts in question are only relevant to the agent's freedom *to the extent that* they help ground the relevant facts about the agent and her connection to the world. For, on this view, these are the facts in which the agent's freedom is most directly grounded.

The fact that counterfactuals can end up being relevant to an agent's freedom, even on a view of freedom based on actual causes (like mine), shouldn't be a surprise. After all, it's not that uncommon to think that causation *itself*, including ordinary causation involving positive events, is partly grounded in counterfactual facts (according to counterfactual views of causation, this is in fact what happens).[4] So, on the view that I'm proposing, the "ultimate" grounds of freedom will be determined, in the end, by what the grounds of causation are.

Now, one could, of course, be interested in finding out what those ultimate grounds of freedom are. That's a different project, one that requires taking a stand on some metaphysical concepts, such as causation, that help ground freedom. Again, I prefer to remain neutral (or as neutral as possible) on these matters, since doing so results in a more ecumenical view that more people could in principle find attractive.[5]

2 Reply to Mele

In the part of Chapter 5 that deals with the manipulation argument for incompatibilism, my main goal was to respond to this version of the argument:

[4] In turn, notice that it's also not uncommon to think that counterfactual facts are themselves grounded in actual facts—this is "actualism" about modality; a view according to which modality is ultimately reducible to what's actual.

[5] I also discuss this point in Sartorio (forthcoming).

Diana Argument (Simple Version):

(1) Ernie is not morally responsible for his murdering act in the Diana case.[6]
(2) Ernie meets all of the standard compatibilist conditions when he performs that act.
(3) Therefore, all standard forms of compatibilism fail.

My strategy was to respond to that argument by attacking the reliability of the intuition behind (1). Whereas I granted that (1) seems intuitively true, I argued that we should *mistrust* that intuition. If we cannot trust that intuition, the Diana argument cannot get off the ground.

My argument that we should mistrust our intuition about the Diana case relied on the fact that the following two claims are true, at least of many of us:

(a) Our reaction to the Lightning Strike case (where Diana is replaced by lightning, a natural phenomenon, but everything else is kept the same) is significantly different from our reaction to the Diana case.
(b) It is clear to us that Ernie's responsibility is the same in Diana and Lightning Strike (we regard the mere presence of an agent instead of a natural phenomenon, by itself, as obviously irrelevant to Ernie's responsibility). Let's call this claim *the irrelevance thought.*

As Mele notes, I'm a "troubled compatibilist": I think the irrelevance thought is clearly true, but I still have diametrically opposed reactions to the two cases—I have the intuition that Ernie is responsible in the Lightning Strike case and not responsible in the Diana case. Now, Mele rightly points out that the situation is likely to be different with an *agnostic*, "Agnes": Agnes is likely to share the intuition that Ernie is not responsible in the Diana case but feel genuinely *undecided* about the Lightning Strike case; that is to say, she is likely to lack any clear intuition whatsoever about that case. Mele suggests that my response to the Diana argument won't be persuasive for agnostics like Agnes. The worry, I take it, is that the non-responsibility intuition about the Diana scenario can have more force in the case of Agnes, for Agnes could use that intuition, together with the irrelevance thought, as a reason to *resolve her previous indecision* about the Lightning Strike case in favor of a non-responsibility judgment.

Should my argument persuade an agnostic like Agnes? I think it should—at least, again, assuming she meets both of the aforementioned conditions, (a) and (b). If Agnes is just as committed to the irrelevance thought, but she also finds herself having significantly different reactions to the two cases (in particular, in her case, an intuition of "not responsible" in the Diana scenario versus the lack of any clear intuition in the Lightning Strike scenario), I think this should lead her, too, to be skeptical of the non-responsibility intuition in the Diana case. In other words, she should withhold judgment about *both* scenarios.

[6] For a description of the Diana case see Mele's contribution in this issue.

Consider the following analogy. Imagine that we are testing our intuitions about the responsibility of an agent, Bernie, in two scenarios. In both scenarios Bernie makes a morally significant choice under very similar circumstances. The only difference between the two scenarios is the presence of a black cat: in one case there is a black cat in the background and in the other case there isn't. When the cat is around, he doesn't affect Bernie's deliberation in any way, so his presence is clearly irrelevant to Bernie's responsibility for his choice. Still, imagine that we find that people's reactions vary significantly depending on whether they are told that the black cat is in the background. Imagine, in particular, that we find that, when the description of the scenario includes a black cat, most of us see Bernie as not responsible for his choice, but we report being unsure or undecided when the description of the scenario doesn't include the cat. If this is what we find, we'll probably suspect that something fishy is going on (that something is messing with our intuitions). We certainly *won't* see the non-responsibility judgment in the black cat scenario as a reason to resolve our previous indecision about the cat-free scenario in favor of a non-responsibility judgment. Instead, we'll just withhold judgment altogether.

Similarly, it seems to me that, to the extent that Agnes is deeply committed to the irrelevance thought, but still finds herself having significantly different reactions about the two scenarios, she should withhold judgment altogether, at least as a first measure. Of course, the same reasoning suggests that, if we had started off with an intuition of *responsibility* in the Lightning Strike case (as a compatibilist might have), we shouldn't trust that intuition either. Basically, in these circumstances we can't trust *any* intuition we may have about the agent's responsibility; all we can do is suspend judgment.

It may be pointed out that, if the only rational thing for Agnes to do is to withhold judgment about both cases, this in fact amounts to favoring the initial position that Agnes has about the Lightning Strike case over the one she has about the Diana case. So, it may be objected, why think that she should suspend judgment altogether, if another way to avoid the inconsistency would have been to embrace a non-responsibility judgment about both cases? Am I not being unfair to the incompatibilist in claiming that Agnes should withhold judgment about both cases, in these circumstances?

In response, let's bear in mind that suspending judgment is a special kind of cognitive attitude: it's not an attitude that consists in believing or disbelieving a certain proposition, but in *failing to* believe or disbelieve it, or in being "committedly neutral" about it. Oftentimes, when we're given reason to think that something has gone wrong with our cognitive processes or the way we evaluate propositions, the rational thing to do is just to suspend judgment, to refrain from holding any particular belief, and thus to avoid incurring any cognitive risks. In other words, suspension of judgment is regarded as the safe, "retreat" state in these cases. In the black cat scenario, for example, we have reason to think that something went seriously wrong and, as a result, we feel like we should suspend judgment altogether. That seems like *the only* rational thing to do in the circumstances. So, unless we have any reason to believe that Agnes's situation is any different, this suggests that Agnes, too, should suspend judgment.

At this point I expect that proponents of the Diana argument will try to point to some significant difference between the black cat scenario and the Diana scenario. In particular, they may suggest that the non-responsibility intuition that we get in the Diana case is *more reliable* because it's somehow revelatory of some genuine threat to our free will that mirrors the threat posed by determinism, which is not something that we see in the black cat scenario. They may insist that Diana is not a "mere distraction" like the black cat, but a useful "proxy," in that thinking about Diana helps us identify and latch onto the threat that determinism itself poses to our free will. As a result, we should give more credit to our intuition about the Diana case; in particular, we should regard it as more reliable than our intuition about the black cat case.

However, note that this would constitute a further move in the dialectic. One would then have to examine the reasons given by the incompatibilist for distinguishing Diana from the black cat. Is there any good reason to think that our reaction to the presence of Diana tracks something important and deep about the agent's responsibility, whereas our reaction to the presence of the black cat does not? The way I see it, the burden of proof now lies on the incompatibilist side. The incompatibilist has to show that Diana is *not* a mere distraction like the black cat (the compatibilist doesn't have to show that Diana *is* a mere distraction like the black cat). So this is one possible way in which I see the dialectic unfolding from here: the incompatibilist could try to draw attention to some potential reason for thinking that the intuition about the Diana case is reliable, or more reliable than that about the black cat case, and the compatibilist would then have to respond to that move. But the compatibilist needn't budge until the incompatibilist makes that further move.

Finally, let me turn to a different aspect of Mele's comments. Mele suggests that my response to the manipulation argument also won't be persuasive for a certain kind of *compatibilist*: a compatibilist who rejects the irrelevance thought. This is someone who believes that, whereas Ernie is responsible in the Lightning Strike scenario, he is not responsible in the Diana scenario (because Diana's presence somehow results in the violation of a historical compatibilist condition for responsibility).

In response, I agree that my argument won't persuade a compatibilist of this kind. Fortunately, I don't need to persuade a compatibilist of this kind. I am only trying to respond to an argument for incompatibilism, and those compatibilists who reject the irrelevance thought believe that they have a good response to that argument, so they don't need to be persuaded that the argument fails. (Of course, those compatibilists and I disagree about what the best response to the manipulation argument is. But my goal here is not to persuade everybody that my response to the manipulation argument is the best; my goal is only to persuade *those who need persuading* that the manipulation argument fails.)

Still, you might think that there is a potential problem here, in that the mere existence of such compatibilists suggests that the irrelevance thought is not as obvious as I made it seem. However, I don't think this would be an accurate representation of what is going on in this case. The existence of those compatibilist views by itself doesn't speak against the plausibility of the irrelevance thought. The

irrelevance thought is extremely plausible, especially given that we are focusing on scenarios that differ only with respect to facts that obtain before the agent was even conceived. Of course, there might be extraordinary circumstances in which even extremely plausible claims have to be given up. But this should only be used as a measure of last resort, and this is, in fact, how I see the compatibilist move that consists in rejecting the irrelevance thought: as a measure of last resort.

3 Reply to Pereboom

Pereboom discusses three main issues in his comments. First, he focuses on the use I make of absence causation in my account of reasons-sensitivity. In Chapter 4 I claim that being sensitive to reasons is a matter of causally responding to absences of reasons of a certain kind. Now, absence causation is a controversial notion. Thus, given that my goal is to be as ecumenical as possible, Pereboom has a helpful suggestion to offer: he suggests that I cast my account instead in terms of the notion of causal explanation, which he regards as broader than causation itself. Pereboom thinks that, whereas the claim that absence causation is controversial, the claim that absences can be part of causal explanations isn't (for example, he thinks that it is widely accepted that the lack of water causally explains the death of plants).

I'm happy to embrace this suggestion, assuming that in thinking of causal explanations we have in mind a notion that preserves some of the main connotations of causation and, in particular, the link to moral responsibility. Helen Beebee (2004) proposes an account of causal explanation that I think would do the job. Like Dowe and others, Beebee rejects the possibility of absence causation, but she suggests that absences can enter in causal explanations, and in roughly the way I think Pereboom is imagining. Imagine that Flora fails to water her plant and the plant dies. On Beebee's account, Flora's omission causally explains the plant's death. Following David Lewis (1986), Beebee suggests that to causally explain an event is, roughly, to give information about its causal history. Lewis noted that it's possible to give information about an event's causal history without identifying any of its specific causes. For example, the claim that JFK died *because somebody shot him* gives information about the causal history of JFK's death (namely, that it contains some shooting event) without specifying a particular shooting event (such as one involving a particular assassin) as a cause. Following up on this point, Beebee suggests that omissions can provide information about the causal history of events, and thus they can causally explain without themselves doing any causing. Importantly, Beebee suggests that the information that omissions can provide is partly modal, in that it also includes information about the causal structure of some nearby possible worlds. To say, for example, that the plant died because Flora didn't water it is to say, roughly, that, in the actual world, no watering event by Flora figures in the causal history of the plant's death, and also that in relevant possible worlds where Flora waters the plant, the plant continues to live as a result (the watering causes the plant's survival). On Beebee's view, all this information is explanatorily relevant to the actual event of the plant's death.

I think that an account of this kind would serve my purposes. Note that Beebee's account bears some important similarities to Dowe's quasi-causation account (my main example in the book of a view that avoids the commitment to absence causation while still preserving many of the connotations of causation, including the link to moral responsibility). In particular, note that both Dowe and Beebee appeal to causal relations in other possible scenarios to ground their claims about quasi-causation or causal explanation. I think this is no coincidence, and that what they are trying to capture with this is a certain kind of connection between the agent and the outcome that can ground the agent's moral responsibility for the outcome, even if that connection is not actual causation. As I pointed out in my reply to Clarke, this is really all I need for the account to work: a way of understanding the relation between agents and the world that can ground their moral responsibility. So I agree with Pereboom that my account could be framed in terms of a view of this kind, and I am grateful, again, for the helpful suggestion.

The second issue discussed by Pereboom concerns a certain kind of disposition: a disposition to respond to reasons and absences of reasons. It's plausible to think that an agent is sensitive to reasons when her behavior is caused by an exercise of a disposition of that kind. But Pereboom thinks that, given the way I have formulated my view, which is centered in the role played by the reasons and absences of reasons, I'm not leaving room to be played by the dispositions *themselves*—unless, that is, we understand the claim that the dispositions cause the behavior as reducible to the claim that the reasons and the absences of reasons cause the behavior. But this is a reductivist view of dispositions that Pereboom doesn't accept.

In response, I don't take a stand on the nature of dispositions, including their reducibility or irreducibility, and I don't think I need to. On my view, the fact that the disposition caused the behavior needn't be reducible to the fact that the reasons and absences of reasons caused the behavior. Imagine that one instead believes that the fact that the reasons and absences of reasons caused the behavior is *grounded in* the fact that the relevant disposition was exercised on that occasion, without one fact being reducible to the other. My view is compatible with this claim. This came up briefly in my reply to Clarke, but it's worth re-emphasizing here: on my view, anything that helps ground the relevant facts about actual causes can indirectly help ground the facts about the freedom of agents. Thus, even if the facts about the role played by the dispositions were not reducible to facts about the role played by the reasons and absences of reasons, they could still ground the freedom of agents in this other way.

Finally, Pereboom turns to my discussion of the manipulation argument, which he regards as a central motivation for incompatibilism. Pereboom starts by making some helpful remarks about the dialectic. He points out, rightly, that different arguments will move different people, depending on their initial reactions about the cases. As I suspected, he doesn't share my intuition about the Lightning Strike case (the natural variation on the Diana case), and, in particular, he reports that his intuitions don't change significantly when we move from Diana to Lightning Strike, as I said I expected many people's intuitions to shift. Still, my hope is that my reply to the manipulation argument can have force for others who are not committed

incompatibilists (where this is supposed to include not just committed compatibilists, but also, hopefully, many agnostics).

Then Pereboom discusses a different case involving a bacterial infection, a scenario where subjects have reported having non-responsibility intuitions similar to those elicited by the Diana case. In this alternative scenario the "manipulation" is not done by an intentional agent but by bacteria. The effect of the bacteria is that the agent's dispositions slowly but increasingly become more egoistic, and this is done, allegedly, without bypassing the agent's agential capacities; still, if Gunnar Björnsson's findings[7] are correct, people tend to have the intuition that the agent is not responsible in this case. So, could one use this "natural" scenario instead of the Diana case to build a manipulation argument against compatibilism? This would allow the incompatibilist to get around my reply to the manipulation argument. For my reply relied on a marked difference in our intuitions between cases involving other agents who could be blamed for the same thing, on the one hand, and natural scenarios where there are no such agents, on the other.[8]

However, note that the bacteria case is a more "localized" case than the Diana case. In the Diana case the manipulation is done before Ernie is even born; here it is done at a later stage. My reason for focusing on the Diana case was the salient possibility of historical compatibilism: forms of compatibilism that allow for a role played by facts concerning the causal history of the agent's behavior. I myself tried to stay neutral about whether the best form of compatibilism is historical or non-historical. But it seems clear that, the farther back one goes, the less plausible it is to suggest that the agent fails to meet some relevant historical condition. So in choosing the more remote cases I was trying to build the strongest possible manipulation argument for incompatibilism.

At the same time, I think it's telling that, when you move far back enough to a time before the agent was born, as in the Lightning Strike case, it's much harder to build a natural case that elicits the same non-responsibility intuition (at least in subjects who are not committed incompatibilists). For then it's natural to ask: why is it that much harder, *if not* because of the absence of another agent like Diana, whose presence would be messing with our intuitions?

Now, at this point Pereboom has a suggestion to offer that could help the incompatibilist. It's a Strawsonian explanation that appeals to the naturalness of our reactive attitudes. On this view, even if (as the incompatibilist would say) the human practice of holding responsible fails to track the truth, it is widespread and it is resilient, in that it is sustained by the strength of our reactive attitudes. Pereboom thinks that this is supported by the observation that shared human practices in

[7] Referenced in Pereboom's contribution in this issue and in Björnsson and Pereboom (2016).

[8] It would also mean that we couldn't use the diffusion of responsibility effect that I used as a debunking explanation of our intuitions about the Diana case. But I take this to be less of a problem for me. As I explained in Chapter 5, the debunking explanation that I provided for the Diana case was merely intended as an exploratory thought. The way I see the dialectic, compatibilists don't actually need to explain why or how the non-responsibility intuition about that kind of case arises. As I noted in my response to Mele, all I think the compatibilist needs to do is provide reason to doubt one of the premises in (the best formulation of) the manipulation argument; offering a debunking explanation isn't needed until the burden of proof has been shifted back to the compatibilist.

general tend to be sustained by strong emotions. So Pereboom would probably suggest that these Strawsonian considerations explain why Ernie appears to be more responsible in Lightning Strike than in the Diana scenario, even though he is not responsible in either scenario. For, whereas Ernie is the only one who could be blamed for his actions in Lightning Strike, Diana can be blamed for them in the Diana scenario.

This is, indeed, a possible explanation of the difference in intuitive judgments between Lightning Strike and Diana, and one that a proponent of the manipulation argument could embrace. Still, why think that it's the best or most reasonable explanation? Again, the way I see the dialectic concerning the manipulation argument, incompatibilists are at a disadvantage (so to speak) in that they are the ones arguing for incompatibilism in this instance. It's *on them* to show that the Strawsonian explanation is the best explanation of the difference between Lightning Strike and Diana; it's not on compatibilists to show that it isn't, for compatibilists don't have the burden of proof.

Of course, it's possible to conceive of scenarios where the burden of proof would be shifted back to the compatibilist. In particular, as I noted in my reply to Mele, I think incompatibilists would succeed in shifting the burden of proof to the compatibilist side if they could somehow show that the intuition about the Diana case is particularly reliable, and that it should be trusted *despite* our reaction to the Lightning Strike case. But, until that happens, the compatibilist shouldn't budge. So, again, this is how I see one way in which the debate could evolve from here. I look forward to continuing this debate with Pereboom and other incompatibilists.

Acknowledgments Thanks to Randy Clarke, Juan Comesaña, Michael McKenna, Al Mele, and Derk Pereboom for comments on a draft of these replies. I am very grateful to my critics for taking their time to think about the ideas in the book and for their insightful comments.

References

Beebee, H. (2004). Causing and Nothingness. In H. Collins & L. Paul (Eds.), *Causation and counterfactuals* (pp. 291–308). Cambridge: MIT Press.

Björnsson, G., & Pereboom, D. (2016). Traditional and experimental approaches to free will and moral responsibility. In J. Sytsma & W. Buckwalter (Eds.), *The Blackwell companion to experimental philosophy* (pp. 142–157). Hoboken: Wiley Blackwell.

Dowe, P. (2000). *Physical causation*. Cambridge: Cambridge University Press.

Lewis, D. (1986). Causal explanation. In D. Lewis (Ed.), *Philosophical papers II* (pp. 214–240). Oxford: Oxford University Press.

Lewis, D. (2004). Void and object. In H. Collins & L. Paul (Eds.), *Causation and counterfactuals* (pp. 277–290). Cambridge: MIT Press.

Sartorio, C. (Forthcoming). Replies to critics. Symposium on *Causation and Free Will, Teorema.*