Criminal Law and Philosophy (2019) 13:551–559
https://doi.org/10.1007/s11572-018-9481-x

BOOK REVIEW

Review of Carolina Sartorio's *Causation and Free Will*

Alex Kaiserman[1] · **Daniel Kodsi**[1]

Published online: 12 October 2018

1 .

The major fault line in the free will literature in recent years has been between *alternative-possibilities* views, which take an agent's freedom to be grounded, at least in part, in whether she could have acted differently than she actually did, and *actual-sequence* views, which take freedom to be fully grounded in the actual sequence of events issuing in the agent's behaviour. In *Causation and Free Will*, Carolina Sartorio defends a novel actual-sequence account, one according to which the actual sequence issuing in an act is just its causal history. What may have appeared like insurmountable challenges to this view can in fact be answered, Sartorio argues, once we properly appreciate various aspects of the metaphysics of causation. In five densely argued but accessible chapters, Sartorio draws connections between previously isolated debates so obvious in retrospect that it's easy to wonder why they weren't noticed before. The result is an elegant account of free action that avoids many of the problems that have plagued its competitors. Thus, while we will question some of Sartorio's specific claims, we welcome the new standards of metaphysical rigour that she has brought to the debate. We recommend her book without hesitation, especially to anyone who still doubts that there is much to learn from the metaphysics of causation about the central concepts of ethical and legal theory.

We start in Sect. 2 with an outline of the major claims in the book, before going on, in Sects. 3 and 4, to raise two potential problems for Sartorio's account of freedom that we believe merit further discussion.

2 .

Sartorio starts, in Chapter 1, by providing some initial motivation for her distinctive 'actual-causal-sequence' theory of freedom. Consider the following familiar case:

✉ Alex Kaiserman
alexander.kaiserman@philosophy.ox.ac.uk

[1] Balliol College, University of Oxford, Oxford OX1 3BJ, UK

> *Neuroscientist:* Frank is deliberating about whether to shoot Furt. Unbeknownst to him, a neuroscientist has been secretly monitoring Frank's brain processes. The neuroscientist can reliably predict the choices that Frank is about to make by looking at the activity in his brain, and can also manipulate Frank's brain in a way that guarantees that Frank will shoot Furt. He plans to intervene if he predicts that Frank will not choose to shoot Furt on his own. As it happens, Frank chooses to shoot Furt on his own, motivated by his own reasons, and without the intervention of the neuroscientist (who correctly predicts that Frank would make that choice on his own). (p. 13)[1]

Intuitively, Frank acted freely in shooting Furt. But, given the presence of the neuroscientist, Frank couldn't have acted otherwise. So, actual-sequence theorists conclude, the alternative-possibilities view is false: freedom is not grounded in the ability to do otherwise.

Sartorio endorses this line of argument. But she also uses it to motivate her own version of the actual-sequence view. Consider the following case:

> *No Neuroscientist*: Frank chooses to shoot Furt, motivated by his own reasons, and does so. This time, there is no neuroscientist monitoring Frank's brain.

Frank's action was clearly free in *No Neuroscientist*, if any action is. But, in all relevant respects, the causal history of Frank's action in *No Neuroscientist* appears to be exactly the same as the causal history of his action in *Neuroscientist*. Someone who thinks that the freedom of an action is grounded in its causal history therefore has an attractive explanation of why, given that the former action is free, the latter must be too. As Sartorio puts her central thesis: "no difference in freedom without a difference in the relevant elements of the causal sequence" (p. 32).

With this background in place, Sartorio goes on, in Chapters 2 and 3, to lay out the structural features causation needs to have in order for her view to work. The first such feature is that *counterfactual dependence is not necessary for causation.* Readers of this journal will be familiar with the 'but-for' test for causation in legal cases, according to which D's act was a cause of V's harm just in case V's harm wouldn't have occurred but for D's act. Yet the but-for test clearly fails in some cases: Oswald's pulling of the trigger caused Kennedy's death, for example, even if there was in fact a second shooter who would have caused Kennedy's death had Oswald decided not to shoot. Sartorio argues that *Neuroscientist* is just such a case—Frank's shooting Furt is caused by the presence of reasons for Frank to shoot Furt, even though the shooting would still have occurred (courtesy of the neuroscientist's intervention) had the reasons not been present.

The second assumption about causation to which Sartorio appeals is that it is an *extrinsic* relation—whether *X* caused *Y* can depend on factors other than the intrinsic properties of the process linking *X* and *Y*. This enables her to respond to a familiar challenge to actual-sequence views based on comparing cases like *Phones* and *No Phones*:

[1] Unless otherwise indicated, all references are to *Causation and Free Will.*

> *Phones*: I witness a man being robbed and beaten. I consider calling the police. I could easily pick up the phone and call them. But I decide against it, out of a combination of fear and laziness.
>
> *No Phones*: Everything is the same as in *Phones* except that, unbeknownst to me, I couldn't have called the police (the phone lines were down at the time). (p. 56)

Intuitively, my failure to call the police was free in *Phones* but not in *No Phones*, even though the intrinsic properties of the actual sequence leading up to my omission are exactly the same in both cases. Although this might be a problem for some actual-sequence accounts of freedom, Sartorio claims that it's not a problem for her view, because the causal history of my omission is not the same in *Phones* as it is in *No Phones*. In *Phones*, my fear and laziness caused my failure to call the police. But, although my fear and laziness caused my failure to *try* to call the police in *No Phones*, Sartorio insists that they didn't cause my failure to call the police, given that the phone lines were down.

In Chapter 4, Sartorio lays the groundwork for her causal account of free will. Some early compatibilists defined freedom as 'doing what one wills.' But as Harry Frankfurt (1971) famously pointed out, an addict might be doing what she wills, even though her actions intuitively aren't free. Frankfurt's response, roughly speaking, was to define freedom as acting on the basis of a desire one desires to have. But this doesn't account for the *willing* addict, who is addicted to, say, coffee, and wouldn't have it any other way. More recent compatibilist views, then, have appealed to the concept of *reasons-sensitivity*: the difference between the addict and the non-addict lies in the fact that the addict is insensitive to the reasons for or against drinking coffee.

A residual question remains, however, about how to cash out the notion of reasons-sensitivity. It is tempting to say that a person is sensitive to some reason R not to act just in case, had R been present, the person would not have acted. But we can imagine cases where a reasons-sensitive person would *not* have behaved otherwise, no matter what reasons were present. Indeed *Neuroscientist* is just such a case. Frank, it would seem, is reasons-sensitive in acting—but had sufficient reason not to shoot Furt been present, he still would have shot Furt (as a result of the neuroscientist's intervention).

Sartorio's response to this problem is ingenious, and coheres nicely with her actual-causal-sequence account of freedom. Most free actions, she argues, are caused not only by the presence of reasons to act, but also by the *absence* of reasons not to act. Among the causes of the non-addict's action, for example, are the absence of anyone offering her £100 not to drink coffee, the absence of dirt in the cup, and so on. This remains true, Sartorio claims, even if a neuroscientist would have intervened to ensure that the non-addict still drank the coffee had those reasons been present. These absences, however, are *not* part of the causal history of the addict's action, according to Sartorio; and that's why the addict's action isn't free.

Sartorio generalises this idea into the following definition of reasons-sensitivity:

> An agent is reasons-sensitive in acting in a certain way when the agent acts on the basis of, perhaps in addition to the *presence* of reasons to act in the relevant way, the *absence* of sufficient reasons to refrain from acting in that way, for an appropriately wide range of such reasons. (p. 132)

'On the basis of' here is intended as a causal locution—roughly speaking, to act on the basis of the presence or absence of a reason is for one's action to be caused, in a suitably non-deviant way, by the presence or absence of that reason.

In the final chapter, Sartorio addresses three kinds of arguments for incompatibilism. Her response to the first two—so-called 'ultimacy arguments' and 'direct arguments'—is that they beg the question against the compatibilist. Consider the following version of the ultimacy argument, for example:

1. If determinism is true, we are never the ultimate sources of our choices.
2. We cannot be free unless we are the ultimate sources of our choices.
   Therefore,
3. If determinism is true, we cannot be free.

Sartorio argues that, on the best interpretation of what it is to be the 'ultimate source' of our choices, (1) is trivially true, and (2) is equivalent to (3)—hence the argument simply assumes that which it purports to establish.

The third kind of incompatibilist argument focuses on cases in which an agent is manipulated into acting by another agent. Intuitively, such actions aren't free, even though they may well satisfy all of the central compatibilist conditions on freedom. In response, Sartorio tries to force a stalemate with the incompatibilist by arguing that, although some manipulation cases do elicit incompatibilist intuitions, others—featuring non-agential manipulation—do not; and therefore our intuitions about neither kind of case can be trusted.

3 .

While there is much about this sharp and resourceful book that merits further discussion, we will restrict ourselves in what remains of this review to raising two potential problems for Sartorio's account. The first concerns an apparent tension between Sartorio's analysis of *Neuroscientist* on the one hand and her analysis of *No Phones* on the other. Intuitively, Frank's shooting of Furt in *Neuroscientist* is free, whereas my failure to call the police in *No Phones* is not free. According to Sartorio, this is because, while Frank's shooting was caused by, say, his dislike of Furt, my failure to call the police wasn't caused by my fear of the police. But these cases have a very similar structure—they both involve an action (or omission) which, because of some extrinsic feature of the case, would still have occurred regardless of the presence of reasons to act (or refrain from acting). In virtue of what, then, are we justified in attributing causation in one case but not the other?

Sartorio is sensitive to this worry. To address it, she appeals to the idea that causes *make a difference*, in the following sense:

> *Difference-Making (Causes)*: Causes make a difference to their effects in that the effects wouldn't have been caused by the absence of their causes. (p. 94)

Sartorio then claims that (i) Frank's reasons to shoot Furt in *Neuroscientist* made a difference, in this sense, to his shooting Furt, but (ii) my fear in *No Phones didn't* make a difference to my failure to call the police.

Unfortunately, we find neither of these claims very convincing. Frank's reasons to shoot Furt made a difference, in Sartorio's sense, to his shooting Furt only if the absence of those reasons wouldn't have caused his shooting. But, had there not been any reasons to shoot Furt, the absence of such reasons would have caused the neuroscientist to intervene, and this in turn would have caused Frank to shoot Furt. So if causation is transitive, then the absence of reasons to shoot Furt would have caused Frank to shoot Furt, from which it follows that (i) is false—the *presence* of reasons to shoot Furt didn't make a difference, in Sartorio's sense, to Frank's shooting.

Sartorio responds to this in part by denying that causation is transitive. The absence of reasons causes the neuroscientist's intervention, which in turn causes the shooting, but, she claims, the absence of reasons doesn't cause the shooting. Denying that transitivity holds in every case isn't enough to explain why it fails in *this* case, however; and other than remarking that "intuitively, this seems like the right thing to say about the Frank and Furt case" (p. 104), Sartorio doesn't elaborate. More importantly, we think that there are ways of filling out the case on which it is fairly clear that the absence of reasons to shoot Furt *would* have been a cause of Frank's shooting Furt. An analogy might help make the point. Sartorio appeals to cases like the following to support her claim that causation is not always transitive:

> *Switch:* A train rushes toward a fork in the tracks. If a switch is flipped, the train will take the left track, and if the switch is left in its original position, the train will take the right track. Further on, the left and the right tracks merge, and just after they meet, a damsel in distress is tied to the tracks (Collins, Hall and Paul 2004, p. 40).

Suzy flips the switch. Her flipping is a cause of the train hurtling down the left track, which is a cause of the damsel's death. Intuitively, however, Suzy's action is not itself a cause of the damsel's death; so transitivity fails. But now imagine that we fill out the details of the case a little. Suppose the right track is considerably longer and less well maintained than the left, so that Suzy's switching the train to the left track makes it significantly more likely that it will reach the damsel. Now, it should seem, Suzy's flipping the switch *does* cause the damsel's death, even if the death would still have occurred had the switch not been flipped. Suzy's action does causal work by making the route to the effect easier—by cancelling possible 'threats' to the damsel's death occurring.[2]

Returning to *Neuroscientist*, then, let's imagine that Frank in fact agonised over whether to shoot Furt, taking several hours to decide. In the end, he decides to shoot, but his deliberation was highly vexed and tortuous; hundreds of possibilities and

[2] For discussion of this phenomenon, see Yablo (2004) and Touborg (2018).

risks ran through his mind. Had there not been any reasons to shoot, however, the neuroscientist would have intervened, so that Frank would have endured none of the anguish and made his decision fairly immediately. The actual causal process leading to Frank's action, in short, was *riskier* than the process that would have obtained had reason to shoot not been present.[3] Here, it seems clear that the absence of reasons to act *would* have been a cause of Frank's action. But then it follows that Frank's reasons to shoot Furt didn't make a difference to his shooting, from which it follows, given *Difference-Making (Causes)*, that the reasons didn't cause Frank's shooting. Indeed, Frank's reasons take on the aspect of an unwelcome presence with regards to the effect, by blocking the easier route to its occurrence. Wouldn't it have been easier for the shooting to occur if the neuroscientist had just had his cue to intervene? Yet our intuitions about freedom remain unaffected: Frank's action is as free in this case as it is in a case where the actual process is less risky than the counterfactual one.

Similar considerations apply to Sartorio's second claim, that my fear of the police *didn't* make a difference to my failure to call the police in *No Phones*. On Sartorio's conception of difference-making, this claim is true only if the absence of fear *would* have caused my failure to call. But clearly it wouldn't have! Had I not been fearful, it would not have been the absence of fear that caused my failure to call the police; rather, it would have been the lack of a functioning phone line.[4] What's more, in the version of events where I am not fearful, my courage makes it necessary for the phone lines to be down for the call not to occur. The presence of my fear, however, cancels the threat of the phone lines being restored. Once again, the route to the effect—here, my failure to call—is *less risky* in actuality; given my fearfulness, the effect does not need to worry, so to speak, that the maintenance crew should arrive and fix the phone lines, thereby making it possible for my call to go through.

So it seems, at the very least, that Sartorio needs to say much more in defence of her causal claims about cases like *Neuroscientist* and *No Phones*; so far, it is hardly clear that our intuitions about freedom are lined up with our intuitions about causation in the way her view predicts they ought to be.

4 .

The second potential problem we want to discuss concerns Sartorio's definition of reasons-sensitivity. Here it is again:

[3] Of course, the effect itself is guaranteed, given the presence of the neuroscientist. But just as a vase can be fragile even if it's protected by a force field that will activate should anyone attempt to break it, so too can a causal process be risky even if a backup process will guarantee the occurrence of the effect should the first process fail.

[4] Sartorio claims that my fear doesn't make a difference to my failure to call the police because my fear and its absence "make intuitively the same contribution to my failure to call the police" (p. 99). But this is not an application of *Difference-Making (Causes)* as stated. Rather, it just seems like another way of saying that both my fear and its absence cause my failure to call the police if either does; but whether this is the case is the very question at issue.

> An agent is reasons-sensitive in acting in a certain way when the agent acts on the basis of, perhaps in addition to the *presence* of reasons to act in the relevant way, the *absence* of sufficient reasons to refrain from acting in that way, for an appropriately wide range of such reasons. (p. 132)

Note that the definition refers to *sufficient* reasons to refrain from acting. To see why this qualification is important, suppose a trustworthy person hands Anne a tasty apple and offers her a billion pounds to eat it. Delighted, Anne eats the apple and claims the billion pounds. Anne's action was clearly free. But, given the offer of a billion pounds, Anne was not sensitive, in Sartorio's sense, to most of the absent reasons not to eat the apple. Recall that one of the reasons the coffee addict isn't free, on Sartorio's view, is that the absence of anyone offering her £100 not to drink coffee is not a cause of her drinking coffee. But the absence of anyone offering Anne £100 not to eat the apple was likewise not a cause of her eating the apple; given the much better offer on the table, Anne was perfectly insensitive to the absence of someone offering her much less. The relevant difference between the addict and Anne, according to Sartorio, is the fact that, whereas an offer of £100 not to drink coffee would have been a sufficient reason for the addict not to drink coffee, the offer of £100 not to eat the apple *wouldn't* have been a sufficient reason for Anne not to eat the apple.

This raises a difficult question, however, about what exactly is meant by 'sufficient reason' in this context. Presumably it isn't intended as a *modal* notion. A sufficient reason for X to φ cannot simply be a reason such that, were it present, X would φ; otherwise the offer of £100 not to drink coffee wouldn't count as a sufficient reason for the addict not to drink coffee. Perhaps, then, the idea is that a sufficient reason for X to φ is a reason such that, were it present, φ-ing would be *the right thing for X to do*, all things considered. But, intuitively, an agent can act freely even if she is strongly disposed to do the wrong thing a lot of the time. Suppose Barry is quite fond of coffee, though not addicted to it. But Barry is also a terrible friend; even if drinking his coffee would have prevented him from saving a friend in need, he would still have drunk it. Moreover, Barry is an insufferable show-off; even if someone had offered him £100 not to drink the coffee, he would still have drunk it, just for the attention. Finally, he has no standards of personal hygiene; had there been dirt in his coffee cup, it wouldn't have put him off. Just like the addict, then, Barry fails to be sensitive to lots of seemingly sufficient reasons for him not to drink the coffee. But isn't his action, unlike the addict's, clearly free?[5]

Sartorio might reply that, insofar as our intuitions are that Barry did act freely, it's because he is sensitive to *enough* sufficient reasons not to drink the coffee to count as having acted freely.[6] After all, her definition only requires free agents to be sensitive to 'an appropriately wide range' of sufficient reasons not to act, not *every*

[5] Note that we can't define 'sufficient reason not to act' as a reason the agent herself *regards* as sufficient, since that lets back in the problem of the willing addict.

[6] One drawback to this response is that it possibly jeopardises Sartorio's ability to develop her view in a way allowing for degrees of freedom. This is a shame, since—just as there are different levels of addiction—it seems to us that, in some sense, there can be different degrees of freedom and responsibility.

such reason. But the problem is that Sartorio's account seems unable to draw what appear to be perfectly coherent distinctions between people with very similar dispositions. Isn't it possible to imagine two people who are insensitive, in Sartorio's sense, to *exactly* the same reasons, even though in one case the insensitivity is the product of addiction whereas in the second case it's simply the product of a peculiar character? And shouldn't it follow on any satisfactory account of freedom that the latter is free and the former is not? If so, this is a problem for Sartorio, since on her view both agents act freely if either does.

Here's another way of getting at the problem. Suppose an army captain decides to abandon her current mission in order to attempt a rescue of a stranded member of her regiment. But suppose also that the captain is so committed to the principle of 'no one left behind' that she would have attempted the rescue even if it had been the wrong thing to do—even if, for example, the risks had been too great, or the probability of success too small, or the current mission too important to abandon.[7] Surely the captain was morally responsible for her decision—if the rescue attempt, and therefore the mission, were to fail, she would be very firmly on the hook. But she is also insensitive to many reasons not to attempt the rescue, in Sartorio's sense. Indeed, she might well be disposed to act in just the same way, in a variety of different scenarios, as someone with a mere compulsion to respond to distress signals, whom we wouldn't regard as free in acting. The difference seems to be that, in the captain's case, her commitment to 'no one left behind' itself serves as "an expression of the moral life" (Williams 1995, p. 46). Her decision and reasoning intimately reflect her character, in a way that the addict's compulsion does not; but Sartorio's account is not fine-grained enough to accommodate this difference.

It's worth noting that other versions of the reasons-sensitivity view seem much better placed to deal with this problem. On Fischer and Ravizza's (1999) approach, for example, to act freely is for one's action to have been produced by a *mechanism* of a *kind* that is responsive to reasons; and it seems perfectly possible for the mechanism that produced one action to be of a kind that is reasons-responsive, and for the mechanism that produced *another* action to be of a kind that is *not* reasons-responsive, even if the token causal histories of those actions contain exactly the same absences of sufficient reasons not to act. Of course, any view that appeals to mechanisms must answer the tricky question of how they should be individuated. Sartorio takes it to be an advantage of her view that it avoids the need to answer these kinds of questions about mechanisms. But the considerations above suggest that it may not be possible to do without them entirely.

[7] Consider equally a very familiar moment in certain popular films, when it is demanded of the hero that he act against his character, but—despite very high stakes—he cannot. The film's point, especially in more realistic instances of the genre, is not necessarily that he has done the right thing; it is that, because of the kind of person he is, he could not have done otherwise (for better or worse). As the Gary Cooper character rather wonderfully says in 'High Noon': "I've got to, that's the whole thing."

References

John Collins, Ned Hall and L.A. Paul, "Counterfactuals and Causation: History, Problems, and Prospects", in John Collins, Ned Hall and L.A. Paul (eds.), *Causation and Counterfactuals* (Cambridge, MA: MIT Press, 2004), pp. 1–59.

John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1999).

Harry Frankfurt, "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68(1) (1971): pp. 5–20.

Carolina Sartorio, *Causation and Free Will* (Oxford: Oxford University Press, 2016).

Caroline Torpe Touborg, "Hasteners and Delayers: Why Rains Don't Cause Fires", *Philosophical Studies* 175(7) (2018): pp. 1557–1576.

Bernard Williams, "Moral incapacity", in *Making Sense of Humanity and Other Philosophical Papers* (Cambridge: Cambridge University Press, 1995), pp. 46–55.

Stephen Yablo, "Advertisement for a Sketch of an Outline of a Prototheory of Causation", in John Collins, Ned Hall and L.A. Paul (eds.), *Causation and Counterfactuals* (Cambridge, MA: MIT Press, 2004), pp. 119–137.