Acta Analytica (2023) 38:689–705
https://doi.org/10.1007/s12136-023-00547-2

Rescuing Mele/Robb-Style Cases

Pablo Rychter[1]

Received: 31 October 2022 / Accepted: 25 January 2023 / Published online: 4 February 2023

Abstract
A good part of the philosophical debate on free will and moral responsibility in the last fifty years has revolved around so-called Frankfurt-style cases. One of the most important milestones in this debate is the case described by Mele and Robb (1998), which was intended to avoid some earlier objections directed at Frankfurt's original argument. However, the success of Mele and Robb's case has been contested by Pereboom (2001), Widerker (2003), and Moya (2003, 2017), among others. The present paper aims to vindicate Mele and Robb's (and Frankfurt's) general argument by describing a variation of their case that overcomes or avoids the objections of those authors.

1 Introduction

A good part of the philosophical debate on free will and moral responsibility in the last fifty years has revolved around so-called Frankfurt-style cases. One of the most important milestones in this debate is the case described by Alfred Mele and David Robb (1998), which was intended to avoid some earlier objections directed at Harry Frankfurt's original argument. However, the success of Mele and Robb's case has been contested by Derk Pereboom (2001), David Widerker (2003), and Carlos Moya (2003, 2017), among others. The present paper aims to vindicate Mele and Robb's (and Frankfurt's) general argument by describing a variation of their case that overcomes or avoids the objections of those authors. The paper is structured as follows. In section 2, I describe the dialectical background against which Mele and Robb presented their case, which is described in section 3. In section 4, I describe the objections raised against this case, focusing on the contributions of Pereboom, Widerker, and Moya. Next, in section 5, I describe my own variation of Mele and Robb's case, and show how it can overcome those objections. Finally, in section 6, I address some potential objections to my own proposal.

✉ Pablo Rychter
pablo.rychter@uv.es

[1] University of Valencia, Valencia, Spain

2 The Frankfurt-Style Cases and the Dilemma Defence

In this section, I describe the general dialectical context of our discussion. I'll start by describing the purpose and central features of the original Frankfurt-style cases, and then move on to discuss a reaction to them that some authors have called the 'dilemma defence'. I pay special attention to the dilemma defence because I take it to constitute the starting point of Mele and Robb's discussion. As we will see, they present their case as a way out of the dilemma, accepting some of the views of the dilemma's defender as constraints on their argument.

A Frankfurt-style case (*FSC*) is a thought experiment intended to be part of an argument against the *principle of alternative possibilities* (*PAP*): the principle that 'a person is morally responsible for what he has done only if he could have done otherwise' (Frankfurt 1969, p. 829). As we can see, in Frankfurt's formulation, *PAP* says that the availability of alternative possibilities is a necessary condition for *moral responsibility*. In what follows, though, I will follow the common practice of reading into *PAP* the further thought that alternative possibilities are a necessary condition for an action to be *free*—more specifically, for it to involve the sort of freedom that is necessary for moral responsibility. My official formulation of *PAP* will thus be as follows:

PAPNecessarily, a person *S* acted freely in doing *A* at *t* only if *S* could have avoided doing *A* at *t*.[1]

PAP deserves discussion by itself, but also because it is a major premise in a very important argument for incompatibilism, i.e. the view that the truth of determinism is incompatible with free will and moral responsibility. (Very roughly, the other premise of the argument is the thought that if determinism is true, then nobody can at any time avoid doing what they do. This premise, together with *PAP*, entails the conclusion that if determinism is true, then no one is free.)[2] Thus showing that *PAP* is false, as *FSCs* are intended to do, has important consequences for the compatibilism-incompatibilism debate.

FSCs are, as I said, thought experiments aimed at refuting *PAP*: they are intended to depict a conceptually possible situation in which an agent *S* does *A*, and about which the following two claims are true:

C1*S* is morally responsible with respect to *A*.

C2*S* cannot avoid doing *A*.

I will use the label 'Frankfurt-style case' (*FSC*) as a generic name for all the different thought experiments intended to do this job. The simplest *FSC* is the one originally described by Frankfurt (1969), which I call the 'Original Frankfurt Case' (*OFC*). With several variations on Frankfurt's presentation, *OFC* goes as follows:

[1] I assume the range of *A*s in this formulation to be fairly wide, so as to include 'ordinary actions' involving bodily movements (like shooting a gun), as well as 'mental actions' (like *deciding* or *choosing* to shoot a gun), and also omissions (like *not* shooting a gun, or *not deciding* to shoot a gun). However, following the usual practice in the literature that I am discussing, I will focus particularly on decisions and their omissions.

[2] See van Inwagen (1983, pp. 83–93) for a useful presentation of this standard argument.

OFCJones decided at t_2 to steal Anne's car as a result of his own deliberation. However, unbeknownst to Jones, Black was monitoring Jones's reasoning leading to his decision. Black was interested in Jones's deciding at t_2 to steal the car and so, if Black had thought that Jones would fail to take this decision, he would have intervened (perhaps using advanced neuro-scientific technology) to cause Jones to make the decision. However, Black preferred not to intervene if it was not necessary and, as it happened, it was not. Jones decided at t_2 to steal the car by his own reasoning.

Concerning such a case, most people have the two intuitions corresponding to *C1* and *C2*: that Jones was morally responsible for his decision, and that he could not avoid taking that decision. And these two intuitions are jointly incompatible with *PAP*: if Jones was morally responsible, he must have had the sort of freedom that is necessary for moral responsibility even if, contrary to what *PAP* requires in such a case, he could not have done otherwise.

A common question about *OFC* is the following 'epistemological worry', as I propose to call it. How did Black *know* what Jones was going to do? How did he *know* whether or not his intervention was necessary? Frankfurt does not say much about this (he only says that 'Black is an excellent judge of such things'). Whether or not the epistemological worry is really pressing, it does seem to have motivated the introduction of a further ingredient in a wide range of common *FSCs*: the idea of a *prior sign* that reliably indicates what Jones is going to decide, and to which Black is sensitive. As an example, consider a case that we may call '*OFC+PS*'. This case is similar to *OFC*, except that in *OFC+PS*, Jones has a disposition to blush two minutes before deciding to steal anything, and so he blushed two minutes before deciding, by his own deliberative process, to steal Anne's car, which in turn let Black know that his intervention was unnecessary. But if Jones had failed to blush, then Black would have known that his intervention was called for. Prior sign cases like *OFC+PS* are perhaps the most commonly discussed cases in the extensive literature on *FSCs*.

The introduction of a prior sign in *FSCs* answers the epistemological worry stated above but also opens the door, in the clearest way at least, to what Widerker and McKenna (2003, p. 8) have called 'the *prior sign dilemma defence*, or just the *dilemma defence*': a remarkable objection to the efficacy of *FSCs* that has attracted a lot of attention in the literature. (In fact, Widerker and McKenna go as far as to claim that 'the debate over the success of the Frankfurt examples is the debate over whether it is possible to get around the powerful point made by the dilemma-defender' (2003, p. 9).) Here is one way of presenting the argument:[3]

Dilemma Defence (*DD*)

P1: In a *FSC* with prior sign, either the occurrence of the prior sign is causally sufficient for *S*'s decision or it is not.
P2: If it is not, then *S* could have decided otherwise.

[3] Kane (1996) and Ginet (1996) have presented alternative versions of the dilemma defence that, unlike *DD*, do not focus on prior signs. Mele and Robb intend to respond to these arguments too. However, for ease of exposition, I focus here on *DD*.

P3: If it is, then *S*'s decision was not free.
C: A *FSC* with prior sign is no counterexample to *PAP*.

P1 seems clearly true. (We can understand 'causally sufficient' in P1 as follows: *A* is causally sufficient for *B* if and only if *A* either deterministically causes *B* or is necessarily associated with a deterministic cause of *B*.) P2 and P3, however, require further discussion. In relation to our above case (*OFC+PS*), P2 can be justified as follows: if the sign is not sufficient for the decision, then it was possible that Jones blushed, Black failed to intervene, and Jones failed to decide to steal the car. So there was the possibility that he decided otherwise. In other words, the following sequence is possible in *OFC+PS*:

Jones blushes > Black does not intervene > Jones does not decide to steal the car.

As for P3, the justification commonly provided is only partial, or relative to the truth of incompatibilism: it is said that *if incompatibilism is true*, then the presence of a deterministic causal connection between blushing and deciding to steal the car implies that the decision is not free. So an incompatibilist would not regard P3 as initially plausible. The conclusion of the argument would then also be conditional on the truth of incompatibilism. But even with this restriction, the conclusion is normally regarded as significant, because incompatibilists are part of the intended audience to which *FSCs* are presented. The fact that *FSCs* can be resisted by them is thus commonly taken to be a serious dialectical drawback.

It is in this context that the significance of Mele and Robb's contribution can be appreciated. In fact, one can see their strategy as an attempt to circumvent *DD* by getting rid of the prior sign. On their view, and even if they express doubts about them, the premises of *DD* may be accepted, together with the conclusion that *FSCs* with a prior sign are unsuccessful. And, as a result, their contribution consists precisely in describing a *FSC without* any prior sign, thus showing that such a sign is, as they say, 'an inessential feature' of *FSCs* (Mele and Robb 1998, p. 102 n.). One may conclude on the basis of Mele and Robb's discussion, then, that the proponent of *FSC*s should rather dispense with the idea of a prior sign. This conclusion is on my view basically correct, and I will follow its recommendation later on. But let us first take a closer look at Mele and Robb's own case.

3 The Mele/Robb Case (*MRC*)

Like any *FSC*, the case described by Mele and Robb, henceforth *MRC*, is designed so as to satisfy the two claims already considered above:

C1S is morally responsible with respect to *A*.

C2S cannot avoid doing *A*.

But, in addition, there is a third claim that Mele and Robb adopt as a desideratum for any successful *FSC*:

C3S's doing *A* is not the result of a deterministic causal process.

As I already pointed out in my discussion of the dilemma defence, if *C3* is not satisfied by a *FSC*, then the case loses some of its dialectical force against the incompatibilist: if *S*'s doing *A* were the result of a deterministic process, then the incompatibilist may remain unconvinced that *S* is free in doing *A*. Thus, as we are about to see, the world partially described by *MRC* is a world in which determinism is not true. And in particular, the agent's decision in *MRC* is the result of an *indeterministic* causal process.

In order to describe *MRC*, I propose that we distinguish two layers, as follows:

> ***MRC*** **First layer**: Like Jones in *OFC*, Bob decided at t_2 to steal Anne's car as a result of his own deliberative process. But unlike what happens in *OFC*, whether this process is deterministic or not is not left open. Here it is explicitly stipulated that the deliberative process causing Bob's decision, process *X*, is indeterministic. Black was interested in Bob's deciding to steal the car at t_2 and so he started, at t_1, a deterministic and unconscious process in Jones's brain, process *P*, which would have caused Bob to decide to steal the car at t_2 if *X* had not. Process *P* is activated long before t_2, but it is designed as a fail-safe mechanism. That is, it causes the decision to steal the car at t_2 *only if* that same decision is not caused at t_2 by some other process. As things actually happened, *X* caused Bob's decision and *preempted P*; that is, *P* did not cause the decision, only *X* did. But given the presence of *P*, Bob could not have failed to decide to steal the car at t_2. As Mele and Robb say, 'any future open to Bob after the initiation of *P* in which he is capable at t_2 of making a decision includes his deciding at t_2 to steal the car' (1998, p. 103).
>
> **Second layer**: Although *X* and *P* actually converged at t_2 (*X* preempts *P* from causing what *X* causes), if they had diverged (i.e., if *P* and not *X* had caused Bob's decision to steal the car), then *P* would have prevented *X* from causing, at t_2, any decision other than to steal the car.[4] As Mele and Robb suggest, *P* does this by 'neutralizing' the 'neural nodes' that Bob would have needed to use in order to decide something other than to steal the car (1998, p. 105). As it turned out, Bob did not use those nodes, but had he tried to use them, he would have failed to do so.

Let us now consider how *MRC* satisfies claims *C1–C3* above. First, it initially seems plausible that Bob is morally responsible for his decision, given that it has been made as a result of his own deliberative process. Process *P* did not play any role in bringing about the decision, and Bob's deliberation went exactly as it would have gone if Black had not been around and *P* had not been started. So it seems that *C1* is satisfied. It also seems that *C2* is satisfied: Bob could not have failed to decide to steal the car. Had he failed to decide this as result of process *X*, he would have done it as result of process *P*. Finally, it also seems that *C3* is satisfied; since

[4] Or no decision at all regarding the car. Although Mele and Robb are not explicit about this, it is plausible to understand their case as one in which *not* deciding at t_2 to steal the car is equivalent to deciding at t_2 not to steal it. Although not deciding to *A* is in general not identical to deciding not to *A*, there are cases in which these two things are equivalent for practical purposes, and *MRC* can easily be pictured as one such case. We may picture Bob as consciously deliberating whether or not to steal the car, and as thinking of t_2 as a deadline for making up his mind, in such a way that he would regard his failure to decide at t_2 to steal the car as a way of deciding not to steal it.

process *X* is indeterministic, Bob's decision is not the result of a deterministic causal process. In sum, *MRC* seems to satisfy the three claims *C1–C3*, and, as a result, it seems to constitute a successful counterexample to *PAP*.

However, we will see in section 4 that the success of *MRC* has been contested by several critics, and I will argue in section 5 that their objections can be met by moving from *MRC* to a variation of it that eschews some of its problematic assumptions. Before moving to that discussion, though, I'd like to flag some other important features of *MRC*.

First, it is worth making explicit that *MRC* dodges the dilemma defence. As I stressed near the end of section 2, this was a central motivation for Mele and Robb. Since *MRC* does not involve any prior sign (Black starts process *P* no matter what), the case passes through the two horns of *DD*, both of which depend on the existence of such a sign.[5]

Second, I want to call attention to the fact that, in *MRC*, process *P* does two different jobs, one in each of the two layers that I have distinguished. In the first layer, *P* makes it the case that Bob cannot fail to decide to steal the car. In the second, it ensures that Bob will not make some other decision that is incompatible with the decision to steal the car. *P* performs the first job by being there as a counterfactual preempted cause (in the sense that it would have caused the decision to steal the car, if *X* had failed to do so). And it performs the second job by depriving Bob of the neural resources that he would have needed in order to decide something other than to steal the car. One might think that the way in which *P* performs the first job is by doing the second job, i.e., that it is by depriving Bob of his neural resources to decide otherwise that *P* would have caused the decision to steal if *X* had not. But it is not necessary to picture *MRC* in this way, and this is not how Mele and Robb describe it. (They present the neutralization story of the second layer as 'an extension' of the story of the first layer (p. 104), which is supposed to make independent sense.)[6]. Therefore I suggest that we treat *P*'s two jobs as independent of each other, as this will be useful in what follows.

Third, I would like to address the following question: is *MRC* a 'blockage case'? The term 'blockage case' seems to have its origin in John Martin Fischer's discussion of David Hunt's work (Fischer, 1999, p. 114; Hunt, 2000).[7] The distinctive feature of Hunt's case is the particular nature of the device that Black uses to make Jones's decision unavoidable: it is a device that actually alters Jones's brain so that the 'neural pathways' that could lead to alternative decisions are 'blocked' or 'neutralized'. (Here Hunt draws inspiration from the famous blocked door scenario described by Locke.) *MRC* does involve this sort of blockage, but I want to stress that it does so only in its second layer: it is only there that *P* neutralizes the causal

[5] As Mele and Robb argue, their case also avoids other versions of the dilemma defence that, as noted before, do not depend on the existence of prior signs. As I said, for ease of exposition, I will leave aside these other versions and focus here on *DD*.

[6] See also the presentation of this case in Robb (2020, sec. 4.3.2), where the neutralization story of the second layer is mentioned as a secondary, bracketable ingredient.

[7] It should be noticed that Fischer does not classify *MRC* as a blockage case, introducing it instead as a close alternative to blockage cases.

paths that would lead Bob to alternative decisions. As far as the first layer is concerned, those causal paths need not be neutralized. So in reply to our initial question, we can say that *MRC* does qualify as a blockage case, but also that blockage plays a relatively small part in it. This will be especially relevant for my arguments in sections 5 and 6.

4 Objections to *MRC*

As mentioned above, several objections have been raised against the view that *MRC* is a successful counterexample to *PAP*. In this section, I present the three of them that strike me as the most pressing. Then, in section 5, I will explain how *MRC* may be modified in order to avoid these objections.

4.1 The Efficacy Problem

The first objection is what Widerker (2003, p. 55) calls ‘the efficacy problem’. The problem is that it is not clear in *MRC* what happens at t_2 to the causal efficacy of the deterministic process *P.* On the one hand, it seems that this process must still be *active* at t_2. Otherwise, how come it would have caused Bob’s decision at t_2 if the indeterministic process *X* had not? But, on the other hand, if *P* is active at t_2, then it must also cause Bob’s decision. But in that case, Bob’s decision would be deterministically caused, and then *C3* would not be true of *MRC*. As Carlos Moya (2017) puts it,

> if P is still active at t2, as it must be in order to prevent Bob from deciding not to steal the car, it is difficult to see how the indeterministic process could prevent P from deterministically causing Bob’s decision at t2, with the upshot that Mele and Robb’s example would fail to avoid, against their intention, the deterministic horn of the dilemma. (2017, pp. 116–117)

As I understand it, this objection concerns what I have called the first layer of *MRC*, and the first of the two jobs that I have distinguished for *P*. The worry is that, given that *P* is activated beforehand and no matter what, and given that it is a deterministic process, there is nothing at t_2 to stop it from following its course and causing Bob’s decision. The claim that *P* is preempted seems prima facie inconsistent with the idea that *P* is both active and deterministic. But perhaps there is some feature of the example that explains the preemption of *P* by *X*? In connection with this, Mele and Robb say that the subject in *MRC* is ‘physically and psychologically so constituted’ that if both an indeterministic and a deterministic process ‘coincide’ at the time of decision, the indeterministic process preempts the deterministic one (pp. 103–104), and that is why *X* preempts *P*. But the present objection precisely casts doubt on whether such constitution is really possible. Its description seems inconsistent, as we have noted, with the idea that *P* is an active deterministic process.[8]

[8] See Mele and Robb (2003, sec. 5) for their own reply to this objection, different from the one that I offer in section 5. Even if their reply works for the objection as I have just presented it, it does not apply to the extension of it that I describe in the next paragraph.

As I said, the present objection, as I understand it, concerns the first layer of *MRC*. But a related worry may also arise with respect to the second layer: if *P* neutralizes all the neural pathways that would lead to deciding not to steal the car, thereby making it *impossible* for Bob to make a decision other than the one he makes, then what disqualifies *P* from being a *cause* of Bob's decision? As Mele and Robb acknowledge (2003, p. 130), given a suitably undemanding conception of causation (such as certain nomic subsumption and counterfactual accounts), *P*'s neutralization of the neural pathways does qualify as a cause of Bob's decision. This consideration pulls in the same direction as the initial objection from Widerker and Moya, and questions whether *MRC* succeeds by Mele and Robb's own standards. In reply, Mele and Robb have rejected the relevant accounts of causation that generate this problem (2003, p. 130). But this is a commitment that we need not make: the solution that I offer in section 5 allows us to remain neutral on these general issues about the metaphysics of causation.

4.2 Is *X* Really Indeterministic?

A related but different worry is that the causal process *X* is not really indeterministic, so *C3* is not true of *MRC* for this reason. Whereas the efficacy problem casted doubt on *C3*, on the grounds that Bob's decision must be caused by *P* (or that it is unclear why it is not), the present objection focuses instead on *X* and the suspicion that it is in fact, and contrary to Mele and Robb's intention, a deterministic process. This suspicion arises because, in virtue of the stipulations made in the second layer of *MRC*, it is impossible for *X* to produce, at t_2, any outcome other than the one that is actually produced (i.e., the decision to steal the car). Any other outcome is ruled out because, as we have seen, process *P* blocks all the causal paths leading to other possible decisions. It does so, recall, by 'neutralizing' all the neural nodes that would be necessary for Bob to make a different decision. But given that *X* cannot fail to produce the decision that it in fact produces, shouldn't we consider it a deterministic process?

Pereboom (2001, p. 18) illustrates this worry by considering two different scenarios in which an 'epicurean atom' falls downward. The atom's movement is initially assumed to be indeterministically caused in the two scenarios: the atom falls down linearly, but it *could* swerve randomly at any point. The difference between the two scenarios is that in one of them, but not in the other, the atom falls frictionlessly within a tube that fits it perfectly well. So there is a clear sense in which, in this scenario, the atom *cannot* swerve: the tube's wall would prevent it from doing so. Can we then say that in the second scenario, the movement of the atom is indeterministically caused? Pereboom thinks that the case is at least unclear: 'Whether this line of argument [leading to the conclusion that the movement is deterministically caused] is plausible is difficult to ascertain, but it is not obviously implausible.' This case is analogous to *MRC* and other 'blockage cases': the tube makes it impossible for the atom to swerve by blocking possible movements, in the same way that *P* makes it impossible for Bob to decide anything other than to steal the car by neutralizing the relevant neural nodes. Pereboom's conclusion is that considerations like this suggest that actual causal histories in blockage cases may be deterministic.

4.3 Sensitivity to Reasons

A different objection to the success of *MRC* is presented by Moya (2003, 2017), and draws on the idea that Bob's decision is not free (and so *C1* is false of this case) because a certain condition on freedom and moral responsibility is not satisfied. The condition in question is that the agent's deliberative mechanism should be sensitive to reasons that would recommend alternative decisions: a kind of sensitivity that is constitutive of the *rational control* that characterizes free actions. Following Fischer and Ravizza (1998), Moya spells out this condition in terms of *weak reasons responsiveness*. The idea is then that an agent's deliberative mechanism is appropriately sensitive to reasons when, 'keeping constant the agent's actual deliberative and decision-making mechanism, there are some possible scenarios, or possible worlds, in which there is a sufficient reason to decide and do otherwise, she recognizes this reason and she decides and does otherwise' (Moya 2003, p. 117). Moya argues that in *MRC*, Bob's deliberative mechanism is not reasons-responsive in this sense. Given that process *P* neutralizes all the neural nodes that would lead to deciding something other than to steal the car, the deliberative mechanism that actually leads Bob to the decision to steal the car would lead him to that same decision even if he had considered significant reasons to decide otherwise. He would have decided to steal the car even if, for instance, he had been told that the car is equipped with an anti-robbery mechanism that almost inevitably kills anyone who attempts to steal it. Even if Bob had considered this reason and fully appreciated its force, he would still have decided to steal the car. And he would have decided this, or so it seems, by virtue of the very same deliberative mechanisms that he actually uses. These mechanisms have been severely damaged by *P*'s neutralizing effect, even if this damage becomes evident only by imagining alternative scenarios in which Bob considers reasons that he does not have in the actual case. As Moya concludes, 'in virtue of blockage, the deliberative structures that lead Bob to his decision are defective and unable to respond to reasons. [… Bob] does not have an appropriate rational control over his deliberation and decision, which seriously undermines the judgment that he is free and morally responsible' (2017, Chapter 4). But then, as we noted above, *C1* is not satisfied and *RMC* is not a successful counterexample to *PAP*.

5 An Easy Way Out for *MRC* Sympathizers

Here is my proposed, simple solution for these three problems: *MRC**. In *MRC**, everything is like in *MRC*, with two exceptions. First, in *MRC**, the deterministic causal process *P* is not *preempted* by *X*. Instead, both *P* and *X* are causes of Bob's decision to steal Anne's car. They are independent causes, and each of them is by itself sufficient for the effect. Thus, unlike *MRC*, *MRC** involves *actual* or *symmetric causal overdetermination*.[9]

[9] In this respect, *MRC** is similar to the case described by Funkhouser (2009, sec. v), which is the clearest precedent for the relatively unexplored idea of using symmetric overdetermination for constructing *FSCs*. (Although see Huoranszki (2017, p. 201) for critical discussion of Funkhouser's view.) But *MRC** differs from Funkhouser's case in other respects (most notably, in the description of the 'counterfactual sequence' where the two processes diverge, as explained in what follows). And my general goals and approach to the dialectics are different from Funkhouser's. In spite of these differences, many of the objections and replies to *MRC** that I discuss in section 6 would apply to Funkhouser's case as well.

Overdetermination of this kind is generally assumed to be possible. In the discussion on mental causation, for instance, it is generally assumed that an action *could* be the effect of both a mental and a distinct physical cause. Although some philosophers in this debate claim that it is implausible that this is what generally or often happens, the *possibility* of such symmetric overdetermination is not in question.[10]

The second point at which *MRC* and *MRC** diverge is this: *MRC** does not include the neutralizing story in the second layer of *MRC*. Instead, let's take to it be part of *MRC** that if *P* and *X* had diverged, Bob would have decided to steal the car and he would also have decided not to steal it. That is, he would have made two contradictory decisions: an admittedly puzzling (but hopefully not impossible) state of mind. Other than these two differences (symmetric overdetermination rather than preemption, and no 'neutralization' or 'blockage') everything is like it is in *MRC*. In particular, also in *MRC**, Bob actually decided to steal Anne's car as a result of his own reasoning. (Although he *also* did it in virtue of *P*.)

I claim that *MRC** is a successful counterexample to *PAP* and that it avoids the objections to *MRC* reviewed above. Let us substantiate this claim. First, *C1* seems true of *MRC**: Bob is morally responsible for his decision, given that he made it as a result of his own deliberative process, which ran independently of Black's provisions. Of course, Black's intervention is another, independent cause of Black's decision. So we can also deem Black responsible for Bob's decision. But Black's responsibility does not deprive Bob of his. (If two shooters shot the victim at the same time, neither of them can claim to be innocent just because the other is guilty!)[11] Second, it seems that *C2* is also true of *MRC**: Bob could not have failed to decide to steal the car. Process *X* could have resulted in a different decision, but process *P* could not have. So regardless of what *X* led to, Bob was bound to decide to steal the car.

So far, we can see that *MRC** satisfies the two initial conditions for being a successful *FSC*. But, as I have noticed, as a result of considering the dilemma defence, Mele and Robb put forward a third condition they wanted their case to satisfy: that Bob's decision was not the result of a deterministic causal process (*C3*). Admittedly, *MRC** does not meet this condition, since in *MRC**, *P* is an actual (rather than a counterfactual, preempted) cause of Bob's decision. I will argue below in section 6 that this is not as problematic as it seems, and that putting forward *C3* as constraint on successful *FSCs* was perhaps an overreaction to the dilemma defence. For the moment, let me note that even if *C3* is not true of *MRC**, the following weaker claim is:

C3* *S*'s doing *A* is the result of an indeterministic causal process.

This claim is true of *MRC**, of course, because *X* is (in addition to *P*) a cause of Bob's action. Moreover, we are assuming that process *X* in *MRC** goes exactly like it goes in *MRC*. So if *X* satisfies, in *MRC*, the constraints that incompatibilists place on free decision-making processes, then so does its counterpart in *MRC**.

[10] For example, see Kim (1998, pp. 44–45). He questions the plausibility of overdetermination in the case of mental causation, but does not question its coherence. See Crisp and Warfield (2001) for a discussion of Kim's arguments that is especially sympathetic to overdetermination. For other, more general defences of overdetermination, see Sider (2003) and Schaffer (2003).

[11] Schaffer (2003, p. 30) makes a similar point.

Next, I claim that *MRC** avoids the objections to *MRC* considered in section 4 above, or that these objections have plausible answers when we move from *MRC* to *MRC**. Let's consider these objections in turn.

First, in relation to the efficacy problem, the suspicions that Widerker and Moya express about *MRC* are confirmed in *MRC**. In *MRC**, the deterministic process *P* does cause Bob's decision. We do not have to struggle to understand how and why *P* is preempted by *X*, because it is not. Thus, the mystery of preemption disappears, and so do any doubts about the coherence of this case. In reply, it may be argued that the proposed cure is worse than the disease. The admission that *P* is causally efficacious deprives *MRC** of a distinctive attractive feature of most *FSCs*: that whatever makes the subject's decision unavoidable is (allegedly) *not* a cause of his decision or action. I'll consider this worry below in section 6.4.

Second, Pereboom's worry that *X* is not really indeterministic arises with respect to *MRC* only because it includes the 'neutralization story' in its second layer. It is only because *P* neutralizes the neutral nodes that would lead to different decisions that *X's* indeterministic nature comes into doubt. But *MRC** does not include the neutralization story of *MRC*, and so the present worry does not apply to it. Moreover, there is no doubt that *X* is indeterministic in *MRC**: it caused Bob's decision to steal the car, but it could have caused a different decision. The presence of *P* does not affect this.

Finally, the fact that *MRC** dispenses with any neutralization story also helps with Moya's worries about sensitivity to reasons. Remember that Moya's reason for thinking that Bob is not sensitive to reasons in *MRC* was that his deliberative mechanism had been damaged by *P*'s neutralization of several neural nodes. It is because of this neutralization of nodes that the mechanism in question is not reasons-responsive. But, again, in *MRC** there is no neutralization of any neural nodes, and so Moya's reasons for thinking that Bob's mechanism has been damaged do not apply to *MRC**. However, one may think that the sort of presence that *P* has in *MRC** is also damaging to Bob's deliberative mechanisms: it seems that, in virtue of *P*, the mechanism that produced Bob's actual decision would also have produced that same decision, even if Bob had considered overwhelming reasons for not stealing the car. Two things can be said in reply here. First, it is true that, in virtue of *P*, Bob would have decided to steal the car even if he had considered powerful reasons not to. But in that case, and unlike what happens in *MRC*, Bob would have decided (also!) *not* to steal the car. There is no reason to suppose that this would not have been the upshot of process *X*. Thus, had Bob considered powerful reasons not to steal the car, he would have made the decision that those reasons recommended (even if he would have also made another, incompatible decision). But then, Bob is actually sensitive to reasons. Second, it is not clear to me that if Bob had considered powerful reasons not steal the car, his decision to steal the car would have been produced *by the same mechanism* that actually produced that decision. What is clearly true is that he would have made the same decision, but it is plausible to think that he would have done so as a result of a different mechanism (associated with process *P* rather than process *X*). So it is plausible to think that Bob's actual deliberative mechanism is not damaged.

Let us sum up the results of this section. *MRC** is a variation of *MRC* that differs from it in two respects; it features symmetric overdetermination rather than preemption, and it does not include the neutralization story that *MRC* features in its second layer. We have seen that *MRC** is a satisfactory counterexample to *PAP* that avoids the objections levelled against *MRC*. In the next section, I discuss some potential objections to my proposal.

6 Some Objections to *MRC** Considered

My claim that *MRC** is a successful counterexample to *PAP* is of course controversial. Let us consider some potential objections[12].

6.1 Incoherent Agency

In *MRC**, Bob decides, at t_2, to steal the car. But had he decided at t_2 not to steal the car, he would also have decided at t_2 to steal the car. Thus, he would have made two inconsistent decisions at the same time. This may seem an implausible and weird feature of this scenario, one that casts doubt on Bob's rationality and moral responsibility. This is how Moya presents the worry in a previous discussion of *MRC**:

We have [in the non-actual possibility that *X* leads Bob to decide not to steal the car] an agent that decides at the same time to steal and not to steal Ann's car. This is an explicit and conscious contradiction in practical reasoning, as harmful and paralysing for rationality as it is in the theoretical field. But it is, nevertheless, a perfectly possible situation given Bob's neurological and mental state, with its two simultaneous processes operating in his brain. But an agent with such an internal structure is far from being cognitive and volitionally normal, and the intuition that Bob is morally responsible of his decision may turn very unstable (Moya 2018, pp. 143).

In reply, it must be admitted that Bob's resulting state of mind in the non-actual sequence of events is puzzling and irrational. But the counterfactual possibility of an irrational decision is a common and familiar feature of *FSCs*: also in *OFC*, Jones would have made an irrational decision—a decision that does not fit his deliberation process—if Black had 'shown his hand'. However, one might think that in *RMC** the situation is worse than in *OFC*, because we have two *contradictory* decisions, and not merely one irrational decision. But I think that contradictory decisions (as well as contradictory beliefs and desires) are not uncommon, much less impossible. Imagine Lois Lane saying to herself: 'Tonight I'll meet Superman rather than Clark Kent', or a billionaire deciding to invest in space flights to Hesperus but not to Phosphorus. Of course, it is impossible to act on these decisions, but the decisions themselves look possible. More generally, I suggest that we understand decisions as mental states or events with propositional content, which are characterizable in terms of their functional role. In that respect, decisions are very much like beliefs and desires.

[12] I am especially grateful to Carlos Moya for raising some of the issues discussed in this section.

And in the same way that one *can* (although arguably *should* not) believe that *p* and also believe that not *p*, one could also decide that *p* and also decide that not *p*. These decisions cannot both be realized, in much the same way that the belief that *p* and the belief that not *p* cannot both be true. But although the content of these decisions or beliefs cannot possibly be satisfied, the decisions and beliefs themselves are possible. To put it differently: as happens with other intentional states and events, when two decisions are inconsistent, they represent states of affairs that are incompossible. But that doesn't mean that the decisions themselves are also incompossible. In sum, the admitted weirdness of what is possible relative to *RMC** renders *RMC** impossible. (And of course, all we need as a counterexample to *PAP* is a conceptual possibility.)[13]

There is, however, a lingering worry in the passage from Moya quoted above: doesn't the possibility of Bob making contradictory decisions weaken the intuition that he is morally responsible? I don't think so. The mere *possibility* that an agent entertains contradictory beliefs or makes contradictory decisions should not cast doubt on the rationality of his actual beliefs and decisions. If it did, then the worry would arise also with respect to *OFC*; for also in this case, it is possible for Jones to make an irrational decision.

6.2 The Presence of Alternatives

Given that in *MRC** there is no 'neutralization of nodes', Bob could have decided not to steal the car. Nothing prevented him from deciding this (in addition to deciding to steal it). Deciding not to steal the car is, then, something that Bob could have done and did not do. But then, one might think, there is an alternative open to him, so he could have done otherwise. Moreover, this alternative is *robust* in the sense of being relevant for attributing moral responsibility. So it may seem that *MRC** is not a counterexample to *PAP*.

My short reply to this worry is that it is not true that Bob could have done otherwise, nor is it true that there were alternative courses of action open to him, given appropriate understandings of 'otherwise' and 'alternative'. Let me elaborate. For any actual course of events that includes a decision *D*, there are countless many *different* counterfactual possibilities, but only some of them are *alternatives* to *D* being made. I have just decided to take a walk in the park. I could also have decided to take a *quick* walk in the park, or to take a walk in the park *while wearing a hat*. None of these different possibilities is one in which I did not decide to take a walk in the park. They are different but not *alternative* possibilities relative to my actual decision. In each of them, I decide to take a walk. Now, coming back to *MRC**, it is true

[13] Admittedly, it may be difficult to *picture* or *imagine* the possibility of an (irrational) agent making inconsistent decisions at the same time. But picturing and imagining is not the only way to conceive a situation in order to argue for its possibility. What Chalmers (2002, p. 149) calls 'negative conceivability' is another way; one that does not require any 'positive' imagining, but only the absence of *a priori* contradiction. And there seems to be no contradiction in the concept of mutually inconsistent decisions taken at the same time.

that it is possible for Bob to decide not to steal the car, but given the presence of *P*, it is not possible for him *not* to decide to steal it. Thus, what is open to Bob is the possibility of deciding to steal it and also deciding not to steal it. This is not a possibility in which Bob does not decide to steal the car, in the same way as the possibility of deciding to take a walk and wear a hat is not a possibility in which one decides not to take a walk. So the possibility that is open to Bob is not an *alternative* to deciding to steal the car, and not a possibility of him doing *otherwise*. In other words, what we need in order to obtain a counterexample to *PAP* is a scenario in which the decision that is actually made by Bob is unavoidable, so that it is not possible for him not to make *that* decision, no matter which *other* things he may have decided as well.[14]

6.3 An Unsatisfactory Reply to the Dilemma Defence

As mentioned above, *MCR** does not meet a desideratum that Mele and Robb set for their case: that Bob's decision is not the result of a deterministic causal process (*C3*). And as we have seen, the motivation for this constraint is to avoid the second horn of *DD*, i.e. the idea that if the subject's decision in the *FSC* is deterministically caused, and if incompatibilism is true, then the decision is not free. Or, in other words, if the subject's decision is deterministically caused, then an incompatibilist will not be convinced that the decision is free. In *MCR*, the desideratum stated in *C3* is allegedly met, because *P* is preempted and does not cause Bob's decision. But in *MCR**, *P* does cause Bob's decision, and as a result, it may seem worse than *MRC* as a reply to *DD*.

I have three things to say in reply. First, I am not sure that *MCR** is really worse than *MRC* at satisfying *C3*, because I am not sure that, in *MRC*, *P* does not cause Bob's decision in virtue of neutralizing the nodes that may lead to an alternative decision. (As I argued above and will come back to shortly, it may be argued that in making Bob's decision inevitable, *P* also causes it.) Second, even if *C3* is false about *MCR** (*P* is a deterministic cause of Bob's decision), there is a truth in the vicinity of *C3*, namely, Bob's decision is (also) caused by an indeterministic process. And this indeterministic process may well satisfy incompatibilists' constraints on free decision-making. Thus, *MCR** goes *some* way towards satisfying *C3* and answering *DD*. But, thirdly and more importantly, I do not think that *C3* is a reasonable constraint on the success of *FSCs*. Or, in other words, I do not think that a proponent of a *FSC* should aim to answer *DD*. My reason is this: a counterexample to a philosophical thesis may be successful even if it does not convince everyone, particularly those who subscribe to that thesis or defend views that clearly depend on it. The ability to convince one's opponent is a virtue, but it cannot be the benchmark for the success of a philosophical argument. Indeed, the present case seems to be one in which our opponent is not especially well-positioned to object that our argument is incompatible with their views. This is because, as noted

[14] Notice, by the way, that this line of reply could also help to deal with the 'flicker defence' against *FSCs*, in a way that sidesteps all talk of 'robustness'. Many of the so-called non-robust alternatives that have been pointed to in the literature (like deciding to steal *after showing the prior sign*, or deciding to steal *as a result of Black's intervention*) are not really *alternative* possibilities in my sense. For they are not situations in which Bob does not decide to steal.

above, *PAP* is a *premise* in a very central argument for incompatibilism. Many incompatibilists are led to their view precisely by their belief in *PAP*. And this puts them in a especially weak position to object to an argument against *PAP on the grounds of their incompatiblism*. Compare: if you believe that a member of the department owns a Nissan only because you believe that Tom (a member of the department) owns a Nissan, then you cannot dismiss evidence that Tom does not own a Nissan on the grounds that you believe that a member of the department owns a Nissan.

More generally, we can say that a counterexample is successful when it is (possibly in conjunction with other auxiliary theses) actually incompatible with the targeted view. If the auxiliary theses are controversial, then the counterexample may have limited dialectical force, even if it actually illustrates the falsity of the targeted view. Thus, a successful *FSC* may fail to do all the dialectical work that it has traditionally been expected to do.[15] This point is also relevant to other negative assessments of *FSCs*. For instance, Ferenc Huoranszki (2017, p. 204) argues that *FSCs* involving overdetermination may at best work as 'illustrations' of the view that there can be responsibility without alternatives, but not as self-standing arguments against *PAP* that would convince one's dialectical opponent. But even granting this point, the modest role of actually illustrating a controversial view should not be neglected. *FSCs*, conceived as illustrations, may still be counterexamples to *PAP*, and may even help us to assess the general views on which they depend. In any case, as already noted, the success of *FSCs* as counterexamples to *PAP* does not depend on their dialectical force.

6.4 The Cure Is Worse Than the Disease

Finally, let us return to the worry raised above in relation to the fact that *P* is fully causally efficacious in *MRC**. As we noted, this deprives *MRC** of the distinctive trick of classical *FSCs*: that whatever makes the subject's decision unavoidable is (allegedly) *not* a cause of his decision or action. And this may seem too high a price to pay. In fact, some authors take the following as a further basic constraint on the success of a *FSC*[16]

C4Whatever makes it unavoidable that *S* does *A* is not a cause of *A*.

C4 is false of *MRC** because, therein, the unavoidability-maker (process *P*) is a cause (albeit not the only one) of Bob's decision.

I agree that, as a result of this, *MRC** is less attractive (certainly less beautiful) than *OFC*, but I do not think that it is less successful as a result. So I am inclined to think that *C4* is not really a constraint on successful counterexamples to *PAP*. And if it is not, then we can take our present discussion to be a further step in a process of debunking allegedly essential features of *FSCs*: Mele and Robb show that prior signs are 'inessential features' of *FSCs*. Hunt similarly argues that it is inessential

[15] For a survey of other, more modest work that *FSCs* could do, see Sartorio (2016). Sartorio's use of the word 'counterexample' is different from mine, however.

[16] See for instance Moya (2017, Chapter 4) and Widerker's IRR assumption (2003, p. 53), which he takes from Frankfurt (1969).

to *FSCs* that the unavoidability-maker be ‘a counterfactual device’.[17] My discussion here similarly suggests that the alleged non-causal nature of the unavoidability-maker is also an inessential feature of *FSCs*. Actual, overdetermining causes are as good as preempted counterfactual causes, or ‘counterfactual devices’, for depriving the subject of alternative possibilities.

7 Conclusion

I have discussed a new *FSC* that is based on the previous ingenious case described by Mele and Robb (1998). We have seen that this case improves on Mele and Robb’s own, in that it is not subject to several objections that have been presented to their case. I have also considered several potential worries about my new case and pointed out how they can be dealt with. Therefore, my central conclusion is that *PAP* is false. But in the course of my discussion, and as part of my argument, I have offered some motivation for other views that may be at odds with common assumptions in mainstream discussions of *FSCs*. Let me finish by emphasizing those views. First, in section 6.2, I argued that the idea of *alternative* possibilities should be understood very restrictedly, in a way that makes it unnecessary to require that alternatives be *robust* (since all alternatives are robust). Second, in section 6.3, I argued that proponents of *FSCs* need not be guided by the aim of avoiding the dilemma defence. On this point, I part company with Mele and Robb, and disagree with Widerker and McKenna’s judgement about the role of *DD* in the success of *FSCs*. Finally, I suggested in section 6.4 that it is inessential to *FSCs* that unavoidability-makers are not causes of the agent’s decision. A case may be successful even if the ‘circumstances that leave no alternative’ for the agent do play a causal role ‘in bringing it about that he does what he does’. These three unorthodox views have been part of my defence of *MRC**. Whereas some readers may regard these views as too unusual to be relied on, I think they are promising starting points for future research.

Acknowledgements I am grateful to the following people for helpful discussion of previous versions of this paper: Carlos Moya, Virginia Ballesteros,Maria Alvarez, an anonymous reviewer for Acta Analytica, and audiences at Valencia and Alicante.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. The research leading to this work has been supported by the Ministry of Science and Innovation (Government of Spain) and the European Union [Grant PID2019-106420GA-100/AEI/10.13039/501100011033].

Declarations

Conflict of Interest The author declares no competing interests.

[17] Hunt (2000, p. 217). Incidentally, Moya (2003, p. 119) takes this claim by Hunt to be especially problematic. But I think both Hunt and Moya overstate the difference between blockage cases and traditional FSCs. In both sorts of cases, there is a sense in which the ‘device’ or, more clearly, the intervention is only counterfactual: in none of them does Black actually cause the agent’s decision. And in both sorts of cases, there is a sense in which the intervention or, more clearly, the device is actual: also in *FSCs* Black’s monitoring activity and dispositions to intervene are up and running. So the difference is, on my view, more apparent than real.

References

Chalmers, D. (2002). Does conceivability entail possibility? In T. S. Gendler & J. Hawthorne (Eds.), Conceivability and possibility (pp. 145–200). Oxford University Press.

Crisp, T. M., & Warfield, T. A. (2001). Kim's master argument. *Noûs, 35*(2), 304–316.

Fischer, J. M. (1999). Recent work on moral responsibility. *Ethics, 110*(1), 93–139.

Fischer, J. M., & Ravizza, M. (1998). Responsibility and control: A theory of moral responsibility (Issue 2). Cambridge University Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*(23), 829–839.

Funkhouser, E. (2009). Frankfurt cases and overdetermination. *Canadian Journal of Philosophy, 39*(3), 341–369. https://doi.org/10.1353/cjp.0.0053

Ginet, C. (1996). In defense of the principle of alternative possibilities: Why I don't find Frankfurt's argument convincing. *Philosophical Perspectives, 10*, 403–417.

Hunt, D. P. (2000). Moral responsibility and unavoidable action. *Philosophical Studies, 97*(2), 195–227.

Huoranszki, F. (2017). Alternative possibilities and causal overdetermination. *Disputatio, 9*(45), 193–217. https://doi.org/10.1515/disp-2017-0004

Kane, R. (1996). The significance of free will. Oxford University Press.

Kim, J. (1998). Mind in a physical world: An essay on the mind–body problem and mental causation (Vol. 75, Issue 291, pp. 131–135). MIT Press.

Mele, A. R., & Robb, D. (2003). Bbs, magnets and seesaws: The metaphysics of Frankfurt-style cases. In D. Widerker & M. McKenna (Eds.), Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities (pp. 107–126). Ashgate.

Mele, A. R., & Robb, D. (1998). Rescuing Frankfurt-style cases. *Philosophical Review, 107*(1), 97–112. https://doi.org/10.2307/2998316

Moya, C. (2003). Blockage cases: No case against PAP. *Crítica, 35*(104), 109–120.

Moya, C. (2017). El libre albedrío: Un estudio filosófico. Ediciones Cátedra.

Moya, C. (2018). Respuestas a los comentaristas. *Quaderns de Filosofia, 5* (1), 127–147.

Pereboom, D. (2001). Living without free will. Cambridge University Press.

Robb, D. (2020). Moral responsibility and the principle of alternative possibilities. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Fall 2020)*. Metaphysics Research Lab https://plato.stanford.edu/archives/fall2020/entries/alternative-possibilities/

Schaffer, J. (2003). Overdetermining causes. *Philosophical Studies, 114*(1–2), 23–45.

Sartorio, C. (2016). Frankfurt-style examples. In M. Griffith, N. Levy, & K. Timpe (Eds.), The Routledge companion to free will (pp. 179–190). Routledge.

van Inwagen, P. (1983). An essay on free will. Oxford University Press.

Widerker, D. (2003). Blameworthiness and Frankfurt's argument against the principle of alternative possibilities. In D. Widerker & M. McKenna (Eds.), Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities (pp. 53–73). Ashgate.

Widerker, D., & McKenna, M. (2003). Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities. Ashgate

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.