Philos Stud (2016) 173:1963–1982
DOI 10.1007/s11098-015-0587-x

In defence of the Four-Case Argument

Benjamin Matheson[1]

Published online: 23 October 2015

Abstract Pereboom's (Living without free will, Cambridge University Press, Cambridge, 2001) Four-Case Argument was once considered to be the most powerful of the manipulation arguments against compatibilism. However, because of Demetriou's (Australas J Philos 88(4):595–617, 2010) response, Pereboom (Free will, agency, and meaning in life, Oxford University Press, Oxford, 2014) has significantly weakened his argument. Manipulation arguments in general have also been challenged by King (Ethics 124(1): 65–83, 2013). In this paper, I argue that the Four-Case Argument resists both these challenges. One upshot is that Pereboom doesn't need weaken his argument. Another is that compatibilists still need a response the Four-Case Argument. And another is that we get a much better understanding of the Four-Case Argument, and of manipulation arguments more generally, than is currently available in the literature.

Keywords Derk Pereboom · Four-Case Argument · Manipulation · Compatibilism · Incompatibilism

1 Introduction

Pereboom's (2001) Four-Case Argument is one of the most influential of the manipulation arguments against compatibilism. Manipulation arguments are grounded by a manipulation case that purports to be a counter-example to the compatibilist conditions on moral responsibility. In these cases an agent is manipulated—often by some nefarious neuroscientists—to perform some action. Given that the agent has been manipulated to perform this action, the incompatibilist

✉ Benjamin Matheson
Benjamin.matheson@gu.se

[1] University of Gothenburg, Gothenburg, Sweden

aims to elicit the intuition that the agent is *not* morally responsible for that action. But if the agent is not morally responsible even though she satisfies the compatibilist conditions, then those conditions are insufficient for moral responsibility.

Rather than simply being a counter-example argument against the compatibilist conditions on moral responsibility, the Four-Case Argument is a combined counter-example and *generalisation* argument. As with any other manipulation arguments, Pereboom starts with an initial manipulation case—one which he claims is a counter-example to *all* the leading compatibilist conditions on moral responsibility. But, unlike other manipulation arguments, he then presents two other analogous manipulation cases. Pereboom claims that there are no relevant differences between these three cases; so if the agent in the first manipulation case is not morally responsible, then neither are the agents in the other two manipulation cases. He then compares these three manipulation cases to a case without manipulation, but in which an agent has been causally determined to act in a type-identical manner to the three manipulated agents. Pereboom claims that there are no relevant differences between the three manipulated agents and the merely causally determined agent. If this claim holds, then it follows that the causally determined agent is also not morally responsible. But if that's true, then it seems that compatibilism is false. In support of this claim, Pereboom makes a novel move. He argues that the *best explanation* for the non-responsibility intuition about the three manipulated agents is that that they have been *causally determined by events external to them*. Given that this is also true of the causally determined agent, it follows that he, too, is intuitively not morally responsible. Hence compatibilism is false.

Some compatibilists (e.g. Fischer 2004; Mele 2005; Baker 2006) have argued that there are *relevant differences* between one or more of the manipulated agents and the causally determined agent. If sound, these arguments would undercut the Four-Case Argument. In response, Pereboom (2005, 2008) has argued his manipulation cases can be appropriately modified so that the manipulated agent satisfies the extra conditions that these compatibilists have argued are necessary for moral responsibility. Other compatibilists, e.g. McKenna (2008), have argued that the so-called 'soft-line' strategy of positing further conditions on moral responsibility is ultimately untenable. As McKenna sees it, compatibilists are eventually going to have to take a 'hard-line' in response to these arguments, and such a hard-line entails having to bite a bullet. In an effort to render this bullet more palatable, McKenna flips the Four-Case Argument on its head. Rather than generalising from Pereboom's first case to his fourth, McKenna claims that compatibilists can generalise from his fourth case to his first case. This fourth case features a merely determined agent, so McKenna claims that the appropriate attitude to start with here is that *it is not clear that the agent is not morally responsible*. Given that there are no relevant differences between the four cases, this attitude generalises to the first case. McKenna contends that his, as I call it, 'reverse generalisation argument', is sufficient to undermine the non-responsibility intuition that gets the Four-Case Argument off the ground in the first place.

Demetriou (2010), however, argues that this unqualified 'hard-line' reply is ineffective. She claims that there are multiple understandings of Pereboom's first case, and the hard-line is only effective with respect to certain of them. In response to these other interpretations, Demetriou proposes a powerful soft-line reply. She argues that in at least Pereboom's first case the manipulated agent is not, in fact, an *agent* when he acts. So she concludes that the Four-Case Argument fails.

Demetriou's reply has been influential. It has led Pereboom (2014) to modify—and I think weaken—his argument significantly. In this paper, I argue that Demetriou's reply is unsuccessful, so Pereboom does not need to modify—and thereby weaken—his argument. I shall argue that there is an interpretation of Pereboom's first case that avoids Demetriou's apparent dilemma. That is, there is a version of Case 1 that is immune to both her soft-line and McKenna's hard-line reply. Since Demetriou claims that there is no metaphysically coherent interpretation of Case 1 that elicits the non-responsibility intuition, all that I need to do to undercut her response is present one such interpretation. Before I consider Demetriou's response, there is a general challenge to manipulation arguments—due to King (2013)—that must be overcome. I argue that the Four-Case Argument has the resources to resist this challenge. I bring these resources into the open by providing a superior explanation of the Four-Case Argument than is currently available in the literature. As we shall see, while the inference to the best explanation is a novel aspect of Pereboom's argument, it is not in fact required. What is required is *fifth*, and as far as I know never explicitly discussed by commentators, case that is pivotal to the success of Pereboom's argument. It is to this task that I now turn.

2 King's challenge

King's challenge is an instance of the 'no generalisation objection'.[1] It states that because the non-responsibility intuition stems from the presence of *manipulation* in a manipulation case, that this intuition will not *generalise* to a mere determination case that (by hypothesis) features no manipulation. Since generalisation of the non-responsibility intuition apparently fails, all manipulation arguments—as a class—fail. Note that the no generalisation objection is similar to a soft-line reply except that the no generalisation objection does not posit further conditions on moral responsibility whereas a soft-line reply does. Because it is an instance of a manipulation argument, for the Four-Case Argument to be plausible it must overcome this challenge. In this section, I argue that the Four-Case Argument can overcome this challenge.

The Four-Case Argument starts like any other manipulation argument: that is, with a manipulation case. Pereboom claims that his first case is a counter-example to the leading compatibilist conditions on moral responsibility. According to Pereboom, these are: Frankfurt's (1971) hierarchical conditions, Fischer and Ravizza's (1998) moderate reasons-responsiveness condition, Hume's (1739/1979:

[1] Versions of this objection have also been defended by Kearns (2012) and Schlosser (2015).

399–411) character condition, and Wallace's (1994) moral sensitivity condition.[2] Pereboom claims that his first case might be enough to 'convince some compatibilists to abandon their position' (2001: 112).[3] But he notes it might not convince them all. The point of the remaining cases, according to Pereboom, is to show that 'an agent's non-responsibility under covert manipulation generalizes to the ordinary situation' (2001: 112), and if non-responsibility generalises from a covertly manipulated individual to a merely determined individual, then this would force an incompatibilist conclusion.

Let's start with Pereboom's first case:

> Case 1: Professor Plum [referred to hereafter as 'Plum1'] was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as possible, given his history. Suppose these neuroscientists 'locally' manipulate him to undertake the process of reasoning by which his desires are brought about and modified—directly producing his every state from moment to moment. The neuroscientists manipulate him by, among other things, pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum is not constrained to act in the sense that he does not act because of an irresistible desire—the neuroscientists do not provide him with an irresistible desire—and he does not think and act contrary to character since he is often manipulated to be rationally egoistic. His effective first-order desire to kill Ms. White conforms to his second-order desires. Plum's reasoning processes exemplifies the various components of moderate reasons-responsiveness. He is receptive to the relevant pattern of reasons, and his reasoning processes would have resulted in different choices in some situations in which the egoistic reasons were otherwise. At the same time, he is not exclusively rationally egoistic since he will typically regulate his behaviour by moral reasons when the egoistic reasons are relatively weak—weaker than they are in the current situation. (Pereboom 2001: 112–113)

Is Plum1 morally responsible for killing White? Plum1 is subject to what I shall call 'sporadic continuous manipulation'. It is sporadic because he is not *always* manipulated ('he is often [that is, not always] manipulated to be rationally

[2] On Frankfurt's view, an individual is morally responsible only if the structure of her will is in order (i.e. her effective first-order desires conform to her higher-order volitions). On Fischer and Ravizza's view, an individual is morally responsible only if her action-producing mechanisms are moderately reasons-responsive (i.e. those mechanisms are regularly receptive to reasons and weakly reactive to them). On Hume's view, an individual is morally responsible only if her actions stem from her character. On Wallace's view, an individual is morally responsible only if she is sensitive to moral reasons. These, of course, are only necessary conditions on moral responsibility. But Pereboom (2001: 111) claims that we can assume that the Plums satisfy all the other non-controversial conditions (at least in this context) on moral responsibility, such as epistemic conditions.

[3] Notably, though, Pereboom doesn't include historicist conditions among these compatibilist conditions. He does, however, argue that Plum1 satisfies these conditions after he sets out his four cases—see Pereboom (2001: 120–122).

egoistic'), and it is continuous because when Plum1 is manipulated the neuroscientists 'directly produce his every state from moment to moment'. As a result of initiating and regulating his reasoning processes, the neuroscientists directly control Plum1's actions whenever they press their buttons. Given that he kills White because he is under the control of the neuroscientists, it seems that Plum1 is not morally responsible for killing White. Indeed, it seems that Plum1 satisfies the leading compatibilist conditions, which I discussed above.

With this counter-example in hand, Pereboom presents a second case, which he claims is not relevantly different from his first case:

> Case 2: Plum [hereafter 'Plum2'] is like an ordinary human being, except that he was created by neuroscientists, who, although they cannot control him directly, have programmed him to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the result that in the circumstances in which he now finds himself, he is causally determined to undertake the moderately reasons-responsive process and to possess the set of first- and second-order desires that results in his killing Ms. White. He has the general ability to regulate his behavior by moral reasons, but in these circumstances, the egoistic reasons are very powerful, and accordingly he is causally determined to kill for these reasons. Nevertheless, he does not act because of an irresistible desire. (Pereboom 2001: 113–114)

Is Plum2 morally responsible for killing White? Plum2 has been *programmed* to be just as rationally egoistic as Plum1. Given that Plum1 is not morally responsible and that there are supposed to be no relevant differences between Case 1 and Case 2, it seems that Plum2 is not morally responsible. Again, it seems that Plum2 satisfies the leading compatibilist conditions on moral responsibility. Hence, it seems that Case 2 is also a counter-example to those conditions.

Pereboom's next case moves much closer to a case of mere causal determination:

> Case 3: Plum [hereafter 'Plum3'] is an ordinary human being, except that he was determined by the rigorous training practices of his home and community so that he is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1 and 2). His training took place at too early an age for him to have had the ability to prevent or alter the practices that determined his character. In his current circumstances, Plum is thereby caused to undertake the moderately reasons-responsive process and to possess the first- and second-order desires that result in his killing White. He has the general ability to grasp, apply, and regulate his behavior by moral reasons, but in these circumstances, the egoistic reasons are very powerful, and hence the rigorous training practices of his upbringing deterministically result in his act of murder. Nevertheless, he does not act because of an irresistible desire. (Pereboom 2001: 114)

Is Plum3 morally responsible for killing White? Plum3 has been *conditioned* by his parents and community to have the reasoning processes and desires that will eventually lead to him killing White. Conditioning is effectively a sort of programming, except that it is spread out over time. Given this, Pereboom claims that there are no relevant differences between Plum2 and Plum3. And if Plum2 is not morally responsible for killing White, then neither is Plum3. Again, it seems plausible that Plum3 satisfies the leading compatibilist conditions, and hence Case 3 is also a counter-example to those conditions.

Note that Case 3 is much more like a case of mere causal determination than Case 2 is. There are now no neuroscientists and there is no direct intervention into Plum3's brain. Pereboom then removes all manipulation in his final case, a case of mere causal determination:

> Case 4: Physicalist determinism is true, and Plum [hereafter 'Plum4'] is an ordinary human being, generated and raised under normal circumstances, who is often but not exclusively rationally egoistic (exactly as egoistic as in Cases 1–3). Plum's killing of White comes about as a result of his undertaking the moderately reasons-responsive process of deliberation, he exhibits the specified organization of first- and second-order desires, and he does not act because of an irresistible desire. He has the general ability to grasp, apply, and regulate his behavior by moral reasons, but in these circumstances the egoistic reasons are very powerful, and together with background circumstances they deterministically result in his act of murder. (Pereboom 2001: 115)

Is Plum4 morally responsible for killing White? Initially a compatibilist might be sure that he is. But Pereboom claims that there are no relevant differences between Plum3 and Plum4. If that's true, then the compatibilist seemingly cannot resist the claim that Plum4 is not morally responsible, and so cannot resist the conclusion that compatibilism is false. But isn't there clearly a relevant difference between Plums 1–3 and Plum4? Namely that Plums 1–3 have been *manipulated by other agents*, whereas Plum4 has not been.

Pereboom anticipates this objection. Once Pereboom reaches Case 4 he doesn't immediately conclude that compatibilism is false. Rather, he first considers this objection, which he attributes to Lycan (1987/1995). Given that Plums 1–3 are manipulated by other agents and that Plum4 is *merely* causally determined, it seems that a relevant difference between Plums 1–3 and Plum4 is that the former are manipulated by other agents, whereas the latter is not. This seems to be a relevant difference, and one which threatens to undermine the Four-Case Argument. According to Lycan, we need only posit a 'no manipulation by other agents' condition on free will (more on this below). In response, however, Pereboom writes:

> … the claim that this is a relevant difference is implausible. Imagine a further case that is exactly the same as, say, Case 1 or Case 2, except that Plum's states are induced by a machine that is generated spontaneously, without intelligent design. Would he then be morally responsible? The compatibilist might agree that this sort of machine induction is responsibility-undermining as well, and then devise a condition that stipulates that agents are not responsible for actions manipulated by agents or machines. But this move is patently ad hoc. What explanation could there be for the truth of this condition other than that it gets the compatibilist the result he wants? (2001: 115–116)

Pereboom's point is that the compatibilist lacks independent motivation for the claim that manipulation by other agents is a relevant difference between Cases 1–3 and Case 4, and thus a ‘no manipulation’ condition is untenable. If they claimed this was a relevant difference, they would have to make a similar claim about neural state induction by machines. He claims that we can replace the agent-manipulators (that is, the neuroscientists) in Cases 1 and 2 with machines without affecting the intuition that Plums 1 and 2 are not morally responsible (and presumably Pereboom thinks this can be done with Case 3 too). Notice that Pereboom is clearly appealing to at least one further case here—namely one with a machine ‘manipulating’ Plum.

Pereboom then asserts that:

> The *best explanation for the intuition* that Plum is not morally responsible in the first three cases is that his action results from a deterministic causal process that traces back to factors beyond his control. Because Plum is also causally determined in this way in Case 4, we should conclude that here too Plum is not morally responsible for the same reason. More generally, if an action results from a deterministic causal process that traces back to factors beyond the agent’s control, then he is not morally responsible for it. (Pereboom 2001: 116; my emphasis)

This inference to the best explanation might seem a bit confusing. Pereboom writes as if he hasn’t yet shown that compatibilism is false and thus that he needs some further reason to establish that. But, by this point, hasn’t Pereboom, if his argument so far is sound, already shown that compatibilism is false? After all, it seems plausible that there are no relevant differences between Cases 1–3 and Case 4 (assuming that Pereboom is correct that manipulation by other agents isn’t a relevant difference); so it should follow that Plum4 is not morally responsible and hence that compatibilism is false, because there are no grounds for compatibilists to resist the generalisation of the non-responsibility intuition from Cases 1–3 to Case 4. Adding an inference to the best explanation therefore seems unnecessary to show that compatibilism is false.

Although I think the inference to the best explanation is indeed unnecessary, it will be worth exploring why Pereboom takes it to be necessary. The answer lies with the objection that Pereboom discusses quickly before presenting his inference to the best explanation. Although Pereboom attributes this objection to Lycan (1987/1995), Lycan actually says very little on the matter. In response to manipulation cases of Taylor’s (1974), Lycan (1995: 117) says, ‘[w]hat we object to in these cases is precisely that the victim is the puppet of another person—that his or her *choices* are coerced.’ Lycan then claims that we can add a negative condition—what I’ve called a ‘no manipulation’ condition—to our analysis of free will that rules that agents who are the puppets of others are not free, and therefore not morally responsible either. Of course, as Lycan (1987/1995: 117) is well aware, such a condition seems ‘somewhat *ad hoc*’. So, as Pereboom notes above, it seems compatibilists cannot rely on such a condition.

While I agree that a ‘no manipulation’ condition is unpromising, this objection can be pressed *without* claiming that there is such a condition on free will/moral responsibility. This, in effect, turns Lycan’s soft-line reply (i.e. one that posits a new condition on free will/moral responsibility) into what I above called the ‘no generalisation objection’, defended by King (2013). Instead of interpreting Lycan as proposing such a condition, we can interpret him as diagnosing the *source* (or primary cause) of the non-responsibility intuition about manipulated agents—viz. the fact they were manipulated by other agents. If it is the case that this fact is the source of the non-responsibility intuition, we have a prima facie reason *not* to generalise the non-responsibility intuition to Case 4; this effectively stops the Four-Case Argument in its tracks.

When Pereboom responds to this objection he seems worried that being the puppet of another person is a relevant difference that will stop his effort to straightforwardly generalise the non-responsibility intuition from Cases 1–3 to Case 4, and this is likely because he is aware that the fact Plums 1–3 were covertly manipulated by other agents is important in eliciting the intuition that Plums 1–3 are not morally responsible. Why else would Pereboom include them in his cases if he didn't think they were important in eliciting that intuition? Clearly, they are; but to avoid the 'no generalisation' objection it must be that manipulation *by other agents* isn't *essential* to eliciting the non-responsibility intuition.[4] This, therefore, is the point of Pereboom's claim that we can imagine cases just like Cases 1 and 2 except that the neuroscientists have been replaced with an intentionless machine that does all the same work as them. If we agree with Pereboom that Plum1* (an individual just like Plum1 except that his neural states were induced by an intentionless machine rather than a group of agents) is not morally responsible, then the fact that Plums 1–3 were manipulated by other agents isn't in fact essential to eliciting the non-responsibility intuition.

The inference to the best explanation is deployed at this point. Its aim, it seems, is to 'push through' the generalisation of the non-responsibility intuition from Cases 1–3 to Case 4. Given that we (let's assume) judge that Plums 1–3 are not morally responsible, it seems that we do so *because* they were covertly manipulated. This, after all, is the common factor between the three of them. And given that it seems that Plum4 has *not* been manipulated by other agents, we can resist the generalisation of the non-responsibility intuition from Plums 1–3 to Plum4. Pereboom's inference to the best explanation is, in effect, a way to overcome this resistance and to push through the generalisation. If Pereboom is correct that the

[4] Some, following what Todd (2013: 202) says in an effort to rebut Kearns' (2012) version of the no generalisation objection, might hold that manipulation by other agents is essential to eliciting the *intuition* that Plums 1–3 are not morally responsible, but not essential to the *fact* they are not morally responsible. Thus the fact mere determination cases feature no manipulation does not hinder the generalisation of the non-responsibility intuition. Such a move is dubious, however. Suppose you believe that *p* because of evidence *e* but you are then told that *e* is irrelevant to *p*. It seems plausible that you would no longer believe that *p*, since you've been told that your evidence that *p* is *not* constitutive of *p*. Moreover, it seems you *should* no longer believe that *p* if you're evidence that *p* is not constitutive of *p*. In other words, you have no reason to believe that *p* because you lack evidence that *p* is true. But Todd and his followers are committed to saying that you should *and* would continue to believe that *p* even after accepting that your evidence that *p* bears no relation to the truth of *p*, because they are committed to saying that you should and would continue to believe that Plums 1–3 (and other manipulated agents) are not morally responsible even after you are told that the fact they have been manipulated is irrelevant to their non-responsibility, even though the only reason you believe they are not morally responsible is the fact they have been manipulated. Hence this move is at least implausible and perhaps even incoherent. Cf. Schlosser (2015: 82).

best explanation for intuition that Plums 1–3 are not morally responsible is that their actions are the product of a deterministic causal process that traces back to factors beyond their control, then this entails that Plum4 is also not morally responsible as his actions are also the product of such a deterministic causal process.

But there's no reason to deploy the inference to the best explanation here. If Pereboom's extra case (or cases) that feature no agent-manipulators elicit the non-responsibility intuition, then Pereboom has undercut the compatibilist resistance to the generalisation of this intuition to Case 4—that is, the no generalisation objection, as defended by King, can be overcome without the inference to the best explanation. And if Plum4 is not morally responsible, then compatibilism is false. This is not to say that Pereboom can't deploy the inference to the best explanation; it's just to say that he doesn't need to if his goal is show that compatibilism is false.[5] Let us now turn to Kristin Demetriou's response to the Four-Case Argument.

3 Demetriou's reply

Demetriou (2010) argues that the Four-Case Argument fails. She argues that the details of Case 1 are ambiguous and that they are open to several interpretations. She claims that each interpretation can be responded to with either a hard-line reply or a soft-line reply. That is, we can either reject the intuition that Plum1 is not morally responsible by arguing that the intuition is unreliable (by employing McKenna's reverse generalisation argument) *or* claim that there is a relevant difference between Plum1 and a merely determined individual. I shall focus on the interpretation of Case 1 that Demetriou claims avoids a hard-line reply, but not her soft-line reply. My reason for doing so will soon become clear.

While other soft-line replies identify a condition on moral responsibility that Plum1 has not met, Demetriou identifies an alleged condition on *agency* that he has not met. Demetriou points out that a merely determined individual is *causally integrated* in a particular sort of way—that is to say, if we observe the causal structure of a merely determined individual, we will see that his neural (or brain) states realise particular mental states, and these neural states cause further neural

[5] Mickelson (2015) argues that inferences to the best explanation are required by any manipulation argument if it wishes to pinpoint what is freedom- and responsibility-undermining in a manipulation case. That might be so, but an incompatibilist need not pinpoint what is freedom- and responsibility-undermining in a manipulation case to show that compatibilism is false. All she need do is show is that there are no causally determined agents who exercise free will or who are morally responsible for their actions. Contra Mickelson, I think incompatibilism comes with no explanatory burden. It might be that Pereboom, as someone who holds that *libertarian* free will is possible, must explain why determinism is incompatible with free will and moral responsibility in order to establish that libertarian free will is possible (though this isn't obvious). But this only shows that incompatibilists *who are also libertarian-possibilists* have an explanatory burden; it doesn't show that mere incompatibilists do because incompatibilists are only committed to the falsity of compatibilism. And the Four-Case Argument qua argument for incompatibilism need not take on the burdens that Pereboom has in virtue of being a libertarian-possibilist. In other words, while the Four-Case Argument is Pereboom's argument, it has a life independent of Pereboom's views.

states and those mental states cause further mental states.[6] Figure 1 illustrates this causal integration.

Compare this now to Fig. 2, which illustrates Plum1's apparent causal structure.

Plum4 is a counterpart of Plum1 except that Plum4 has not been subject to what Demetriou (2010: 608) calls 'suppressive' manipulation, of which sporadic continuous manipulation is a form. The manipulation is suppressive because it apparently undermines the causal efficacy of Plum1's prior mental states: his subsequent states are caused, not by his previous states, but by the neuroscientists' activities. Plum1's brain and mental states are causally isolated: one brain state realises a single mental state, but his brain and mental states do not cause his subsequent brain and mental states. On the other hand, there are no impediments to Plum4's brain states causing his subsequent brain states, which realise his subsequent mental states. Although Plum1 and Plum4 have exactly the same mental states and therefore have qualitatively identical mental lives, Demetriou claims that they do not share 'the same status in terms of agency' (2010: 608)—that is, while Plum4 is an agent when he kills White, Plum1 is not when 'he' kills her.

Cast in this way, Demetriou's reply is very convincing. She appears to have identified a relevant difference that shows that Case 1 is not a counter-example to the leading compatibilist conditions on moral responsibility, and because she claims to have identified a condition on agency that Plum1 has not satisfied, her response can be used by all compatibilists. But Demetriou's argument supports a rather strong claim. She claims to have shown every metaphysically coherent interpretation of Case 1 succumbs to either a hard-line or soft-line reply. To undermine her argument, then, I need only show that there is *one* metaphysically coherent interpretation of Case 1 that avoids her soft-line reply and that avoids a hard-line reply. I'll start by arguing that there is an interpretation of Case 1 that avoids her soft-lines reply, and then I'll argue that a hard-line reply (i.e. one which uses the reverse generalisation argument) is ineffective in response to this interpretation.

The viability of Demetriou's soft-line reply depends on the plausibility of two claims: (1) that the suppression of an individual's prior states must undermine an individual's agency, and (2) that there is no causal integration *after* the neuroscientists' first intervention (B1/M1 in Fig. 2). I will first argue that (2) is false. While it might be plausible that the neuroscientists' first intervention (i.e. the induction of B1/M1) suppresses Plum1's prior states (i.e. those which precede those shown in the diagram), which I'll grant for the sake of argument, there is a plausible understanding of the manipulation such that there is causal integration after the first intervention (i.e. there is causal process running between B1/M1–B2/M2–B3/M3). I will then argue that (1) is false. I argue that there are lots of everyday instances in real life where an event might 'suppress' prior states *without* an individual's agency being undermined.

The diagrams that are pivotal to Demetriou's argument are normally used to display the relationship between brain states and mental states (see n.7). They do not

[6] How we interpret the details, as Demetriou (2010: 607, n.8) notes, depends on our view about mental causation. What I say in the remainder makes no claims about mental causation, and I remain neutral on this issue throughout.

Fig. 1 (Demetriou 2010: 607) Plum4's causal structure[7]

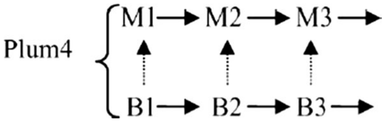
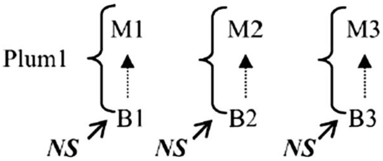


Fig. 2 (Demetriou 2010: 607) Plum1's apparent causal structure



accurately portray the causal structure of the human brain/mind. The causal workings of the human brain are clearly much more complicated that the diagrams suggest.[8] In particular, Demetriou's diagrams suggest that normally a particular brain/mental state is sufficient for the production of a subsequent brain/mental state. But it is not the case that a particular brain/mental states is *alone* sufficient for the production of a subsequent brain/mental state; there are lots of background conditions that must also be satisfied. This is in much the same way that the impact of a billiard ball A is not alone sufficient to move a billiard ball B; it also matters what force A was moved with, what the surface the balls are on is like, what forces (e.g. gravity) are in play, what mass the billiard balls are, and a whole host of other background conditions. The same is true with causation between states. For a particular brain/mental state to produce another brain/mental state, certain background conditions must be satisfied. One of the most important of these background conditions is often the *external world*.

It is uncontroversial that our reasoning processes are regularly influenced by the external world. Perceiving things in the world might cause us to consider certain options or it might help us to make a decision we had been struggling to make. For example, a motivational poster that asserts that we should *just do it* might make us think about the things we can do (when previously we might not have considered these options) or it might help us to decide to do something we had been deliberating about. Such *conscious* influence on our reasoning processes is common. There is also substantial empirical evidence that the world can *un*consciously affect our reasoning processes. For example, prisons have often been painted Baker-Miller Pink in an attempt to reduce aggression among inmates; several studies have

[7] As Demetriou (2010: 607, n.8) notes, these diagrams are usually used to present 'different visions of mental causation'. Also, Demetriou talks about brain rather than neural states as I have been. This difference in terminology is not significant in what follows.

[8] Demetriou (2010: 607, n.10) is aware of this; she claims that her diagrams can 'be seen as scoping down on the precise location of the failure of agency that occurs where the neuroscientists causally regulate the isolated area of Plum1's brain/body which constitutes his reasoning process.' But, as I'll argue, 'scoping down' results in us ignoring details relevant to the issues at hand.

supported the hypothesis that this is effective.[9] Studies have also suggested that smells and sounds can affect our behaviour without us noticing.[10] Hence the conscious and unconscious causal role of the world in the production of actions is something we (and any adequate account of agency) must acknowledge.[11]

As the world has a causal role in the production of action, external events are often important causal contributors—that is, jointly sufficient with our states and other background conditions—in the production of our subsequent states and behaviour. Suppose I feel hungry (B1/M1), so I look in my fridge and I see a cake (B2/M2). I then form the belief that there is a cake in my fridge (B3/M3). As I am hungry and I desire not to be hungry, I make the decision to eat the cake (B4/M4). I then eat the cake. Clearly, there being cake in my fridge is integral to me forming the decision to eat the cake. If there were no cake in the fridge (and I wasn't hallucinating), then I would not have decided to eat the cake. Demetriou's diagrams, however, make no room for the causal influence of the external world. Once we acknowledge this, it becomes plausible that the neuroscientists do not have to undermine the causal efficacy of Plum1's states after their first intervention in order to manipulate him from moment to moment.

Let's think about Case 1 again. According to Demetriou's interpretation of the mechanics of the manipulation, each of the neuroscientists' interventions suppresses a prior state of Plum1's. So when B1/M1 is induced this suppresses whatever Plum1's prior brain/mental state was by rendering those states causally impotent, when B2/M2 is induced this suppresses B1/M1, when B3/M3 is induced this suppresses B2/M2, and so on. In other words, there is no causal process running between B1/M1–B3/M3. But that's a really strange thing for the neuroscientists to do. They, after all, are trying to manipulate Plum1 so that he kills White; they are trying to do so by inducing his brain/mental states so that he engages in an egoistic reasoning process which will lead him to killing White. Now if this is their aim, why would each intervention (after the first intervention) necessarily undermine the causal process between B1/M1–B3/M3? That seems somewhat counter-productive. Presumably the initial state they induce could easily play a role in producing the subsequent states the neuroscientists wish Plum1 to have. The neuroscientists would then only have to intervene to the extent that B1/M1 on its own is not sufficient for the production of B2/M2. Thus, after the first intervention the neuroscientists need only act as a kind of *causal enforcer*—that is, the neuroscientists' activities would *not* be sufficient for B2/M2 to obtain, but their activities would provide whatever extra is required for B2/M2 to obtain. The neuroscientists are, in a sense, causing B2/M2, but they are not causing it alone; B1/M1 just has a more important role to play.

[9] Bennet et al (1991); Profusek and Rainey (1987); Schauss (1985)).

[10] Baron (1997).

[11] The UK government also has a department—affectionately known as the 'nudge squad'—which works on 'improving' people's behaviour in subtle ways. In other words, this department seeks to manipulate people. See https://www.gov.uk/government/organisations/behavioural-insights-team/about for more details.

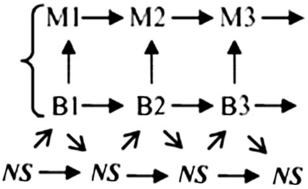
This is similar to how a friend might continue to badger us to do something even after we have agreed to do that something. While there are lots of times our friend's badgering might overdetermine (i.e. we will do the thing regardless), pre-empt (i.e. we would have done the thing, but the badgering is sufficient to make us do it independently), or even be causally irrelevant (i.e. the badgering has no effect on us) to the thing we've agreed to do. There are other times when we want to do something, but we will only actually do it *if* we are badgered. In such cases, the badgering alone is not sufficient for us to act; rather, it is only sufficient in conjunction with our prior states.

After the first intervention, the neuroscientists need only match the causal influence of such a friend in the latter sort of case. The neuroscientists would still be manipulating Plum1 'moment to moment' because they are still controlling exactly what his states are like from moment to moment; it's just that they have less work to do after the first intervention, which does much of the heavy lifting. In the same way, if a friend thinks she has convinced us to do something (and she is smart enough to avoid being causally irrelevant, overdetermining or pre-empting us) she has less work to do to keep us on track. But she might still have some work to do because we might change our minds.

Of course, our friend would only realise she had more work to do if it seemed like we might change our minds. She might be able to judge that this might happen by observing our change in body language. Our body language (among other things), then, acts as a kind of feedback mechanism. So, for the neuroscientists to successfully manipulate Plum1, there must be some sort of feedback mechanism between Plum1's brain and the neuroscientists. That is, the neuroscientists must have some way of knowing whether the state they have just induced will be sufficient (in conjunction with background factors) to produce the subsequent state they wish Plum1 to have. When a prior state is not sufficient, the neuroscientists can, so to speak, make up the difference and help to produce the subsequent state. But, note, the initial state they induce will have a role to play in the production of Plum1's subsequent state. Figure 3 displays this feedback mechanism.

The arrow between the neuroscientists is there to indicate that there is some causal process obtaining between the neuroscientists' activities. The arrow upwards from the neuroscientists to the brain states indicates their interventions into Plum1's brain. The arrow downwards from the brain states to the neuroscientist is the feedback mechanism. This allows them to tailor their subsequent intervention as required. Indeed, the feedback mechanism is required so that the neuroscientists will know when they need to act so as to strengthen Plum's disposition to do what they want him to. And this is indicative of the fact that in the scenario envisioned, Plum's

Fig. 3 The feedback mechanism



brain and mental states causally influence how the neuroscientists intervene. Thus the neuroscientists' interventions are not causally independent of Plum's brain and mental states, as they are on Demetriou's account of Case 1. Once the feedback mechanism is in place, there no longer seems to be any problem with claiming that B1/M1 causes B2/M2 and that B2/M2 causes B3/M3. It is not as if normally brain/mental states *alone* are sufficient to produce subsequent brain/mental states; there are always background conditions which are jointly sufficient with the prior brain/mental state. These are overlooked in discussions about mental causation because our interest is in the relationship between brain and mental states, and not on *how* brain states are produced.

Demetriou (2010: 607, n.9) seems to allude to this sort of diagram. However, she doesn't take seriously the importance of the feedback mechanism. Once we take this seriously, as I've argued, we see that each of the neuroscientists' interventions need not undermine the causal efficacy of Plum1's post-initial-intervention states. Demetriou claims that the feedback mechanism is not important because 'Plum1's states do not, on their own, causally necessitate that the neuroscientists press the buttons that they do. Rather, they press the buttons they do as a causal result of their own, independent reasoning—meaning that they are free to decide, for reasons all their own, which state to cause in Plum1 at any given moment of the manipulation' (2010: 607, n.9). But this line of argument is problematic. In particular, it is ambiguous between different senses of 'free'. Either the neuroscientists are 'free' in a compatibilist sense or a libertarian one. If the neuroscientists have compatibilist freedom then it must be compatible with them being causally determined to act. But it therefore follows that Plum1's states do (or might) causally necessitate that the neuroscientists press the buttons they do and when they do.[12] But if the neuroscientists have some form of libertarian freedom, then Demetriou has granted victory to the incompatibilist by implicitly conceding that the neuroscientists have a different—and better—sort of freedom to the one the compatibilist claims is compatible with the truth of causal determinism. Clearly, compatibilists cannot concede the neuroscientists have libertarian freedom. So compatibilists must accept that the neuroscientists are possibly causally determined, thereby undercutting Demetriou's point.

As I've interpreted Case 1, it sounds similar to an interpretation that Demetriou (2010: 604–605) considers, where the neuroscientists and Plum1 are joint sufficient causes of Plum1's subsequent states. Even in such a case, she claims that the neuroscientists would have to '*undermine* the causal efficacy of Plum1's states in some way' because, presumably, Plum1 is the sort of individual who would act

[12] Most compatibilists accept a necessitarian conception of the laws of nature, so causal determination and causal necessitation can be taken as equivalent. Only the Humean about the laws of nature denies this. See Beebee and Mele (2002) for more on the relevance of our conception of the laws of nature to the free will debate. Importantly, a necessitarian conception of the laws of nature seems to be implicitly presupposed by Pereboom in making his argument against compatibilism. Demetriou could deny this, of course. But she would have to make this thrust of her argument and would have to provide an explicit defence of this view of the laws of nature. Presumably, then, she accepts—at least for the sake of her response to Pereboom—a necessitarian conception of the laws. So we can take causal determination and causal necessitation as equivalent here.

without the neuroscientists' intervention, and hence his states (without intervention) are normally causally sufficient to produce his subsequent states. So, according to Demetriou, the neuroscientists would have to undermine Plum1's states to some extent, but not entirely, so that those states in conjunction with the neuroscientists' input would produce Plum1's subsequent actions. Moreover, she claims that their interventions would be 'faithful' or 'unfaithful'. If they intervene faithfully, then they cause Plum1 to act as he would have acted without their intervention. If they intervene unfaithfully, then they cause Plum1 to act differently. Demetriou claims the faithful interpretation is open to a hard-line reply, and the unfaithful interpretation is open a soft-line reply. While the case I am sketching is more like an unfaithful intervention (which seems necessary to elicit the non-responsibility intuition), there are important differences between the cases.

First, I have accepted for sake of argument that the neuroscientists' *first* intervention entirely undermines the causal efficacy of Plum1's prior states (that is, his states prior to B1/M1), so B1/M1 is produced solely by the neuroscientists' efforts, whereas in Demetriou's joint sufficiency case Plum1 and the neuroscientists are joint sufficient causes of B1/M1—that is, in Demetriou's case, Plum1's prior states are not rendered entirely causally impotent; the neuroscientists simply undermine those states by (somehow) *reducing* their causal efficacy by (somehow) making those states not quite sufficient for the production of subsequent states. Second, I deny that the neuroscientists have to undermine the causal efficacy of Plum1's states *after* the first intervention—that is, I deny that the neuroscientists reduce the causal efficacy of B1/M1, B2/M2, and B3/M3. After all, it seems plausible that the neuroscientists induced B1/M1 because it would have a causal role in the production of Plum1's subsequent states. So the neuroscientists do not need to reduce the causal efficacy of B1/M1 for B2/M2 to come about, nor do they need to reduce the causal efficacy of B2/M2 for B3/M3 to come about. The neuroscientists can, instead, act as causal enforcers, adding whatever is required (if anything is required) for B1/M1 to produce B2/M2, and for B2/M2 to produce B3/M3 in the same way that the world would have to be a particular way for any individual's brain/mental states to cause a subsequent brain/mental state. According to Demetriou, if Plum1's states are not causally efficacious then he does not exert causal control over his actions. But even if B1/M1 is not alone sufficient for B2/M2 to obtain, that doesn't mean that B1/M1 is not causally efficacious.

The important thing to realise, and what I have been trying to draw attention to, is that there is *much* more to the causal story than simply one brain state causing another brain state. If causal control required that brain/mental states *alone* had to be sufficient for subsequent brain/mental states to obtain, then no one would ever exert causal control because there are always background necessary conditions. We usually ignore these background conditions because they are not the focus of our investigation (e.g. when we are investigating mental causation we are only interested in the relationship between brain and mental states). As I argued earlier, the external world (in many different ways) is a background condition, and the external world and our states are jointly sufficient for the production of our subsequent states. But once we include the external world in the causal story, we do not conclude that we are not agents. The world, we might say, is a causal enforcer: it ensures that our states are causally sufficient for the production of our subsequent states. The neuroscientists, then, simply take on some of the role of the world after their first intervention: they ensure that Plum1's states are causally sufficient for the production of his subsequent states. Just as we have no reason to think that the world suppresses our agency, we have no reason to think that the neuroscientists, on the version of Case 1 I have sketched, undermine Plum1's agency after their initial intervention. Hence, even if the neuroscientists' initial intervention suppresses Plum1's prior states, this does not rule out there being a causal process running between Plum1's subsequent states (i.e. between B1/M1–B3/M3).

If there is a causal process between his states, then when Plum1 kills White *he is an agent*. If she wished to maintain that Plum1 is not an agent, Demetriou would have to argue that there is something significant about the initial 'suppression' of Plum1's states prior to the manipulation period. But this claim is dubious. There are lots of instances in real life where our states are 'suppressed', according to Demetriou's definition, to make way for new ones. For instance, suppose I'm thinking about the Four-Case Argument and suddenly I hear my window being smashed. The event of the window being smashed stops me from thinking about the Four-Case Argument and makes me think about my window being smashed. Rather than thinking, 'how is the inference to the best explanation supposed to work?' I am suddenly thinking, 'why is my window being smashed? Is someone trying to break into my house?' Etc. It looks as if the event of the window being smashed *suppresses* my prior states in the sense that Demetriou claims undermines agency. According to her, something suppresses a state if it undermines the causal efficacy of that state. This seems to happen when my window is smashed. If my window had not been smashed, I would have continued thinking about the Four-Case Argument. But as my window has been smashed, my thoughts change. Hence, the causal efficacy of the brain/mental states that underpinned my thinking about the Four-Case Argument are suppressed by the event of my window being smashed. But it is implausible to suppose that hearing my window being smashed undermines my agency because events like this occur regularly in real life. It seems that any plausible account of agency must accommodate such a regular occurrence.

Perhaps there is some significant difference between Plum1 and me which explains why I am an agent and Plum1 is not. But the only difference between us is the method of the causal suppression of our prior states, and that seems like an *ad hoc* difference in this context. In any case, if there were some non-ad hoc difference, incompatibilists could just *change the case* by modifying the mode of manipulation the neuroscientists use to covertly control Plum1; if the problem is that the manipulation is 'inside' the brain, the neuroscientists need only move the manipulation to 'outside' the brain. They might hire a biochemist to devise a poison and a sound technician to devise a series of subsonic sounds to manipulate Plum1. The poison might initiate an egoistic reasoning process and then the subsonic sounds could be applied as a causal enforcer if it ever looks like Plum1 needs some help to keep reasoning egoistically. The neuroscientists could therefore achieve the same effects as radio wave brain manipulation without any worries about Plum1's agency being suppressed.

But incompatibilists don't need to change this case because there isn't a non-ad hoc difference between a state being suppressed by a radio wave and a state being suppressed by a sound (e.g. the sound of a window being smashed). Demetriou's reply only seems convincing because we overlooked the causal role of the world in the production of action and the possibility of a feedback mechanism between Plum1 and the neuroscientists. Once we see the causal role of the world and include a feedback mechanism, we can understand the causal role of the neuroscientists after their first intervention such that they do not undermine Plum1's agency. Hence, it seems that there is an interpretation of Case 1 which avoids Demetriou's soft-line response.

What about a hard-line reply—that is, one that uses the reverse generalisation argument? According to this argument, we must start with our intuitions about merely determined individuals and then generalise them to covertly controlled individuals.[13] For this reply to be successful, it must be the case that we can generalise our intuition about a merely determined individual to Plum1, who is a victim of sporadic continuous manipulation. Even as a committed (hard) compatibilist,[14] I find that my intuition that Plum4 *is* morally responsible does not generalise to Plum1. There is something particularly worrying about the mode of manipulation that Plum1 is a victim of such that I cannot accept he is morally responsible just because he is apparently no different from Plum4. Therefore, Demetriou's overall argument against Case 1 is unsuccessful because there is an interpretation of Case 1 which avoids both her soft-line reply and a hard-line reply.

4 Conclusion

Since King's no generalisation objection and Demetriou's reply fail, the Four-Case Argument still stands. Pereboom, then, does not need to abandon his earlier version of this argument in favour of his newer version. The main difference with Pereboom's (2014: 74–82) newer argument is that it includes a new version of Case 1 (call it 'Case 1_N'), which sidesteps Demetriou's reply. But, as I'll explain below, Case 1_N is much weaker and doesn't have the intuitive pull of his older Case 1. And given that I've shown that Demetriou's reply fails, Pereboom is not forced to retreat to the weaker version of the Four-Case Argument that is grounded by Case 1_N.

Case 1_N features *ego button* manipulation, which was first devised by Shabo (2010: 367). Ego button manipulation involves 'ramping up' an agent's *already existing* egoistic reasoning processes to cause them to reason—and then act—egoistically. In Case 1_N, we are told that Plum1_N 'is frequently egoistic and

[13] While McKenna generalises a neutral attitude from Case 4 to Case 1, I see no reason why compatibilists can't just start with their intuition about Case 4. I don't see how it begs the question against the incompatibilist to start an argument with one's intuition about a case. See Beebee (2002) for more on question-begging. Even so, my point here can be adapted so that the attitude being generalised is a neutral one.

[14] See Matheson (2014). There I argue that victims of 'global' manipulation (i.e. cases where an individual has a new 'moral personality' implanted) are morally responsible.

sometimes strongly so' (Pereboom 2014: 77). The neuroscientists then press their buttons; Plum1_N reasons egoistically and then kills White. The only catch is that Plum1_N *would not* have killed White had the neuroscientists not intervened. While this case might have some pull on incompatibilists, from a compatibilist perspective it is hard to see what is responsibility-undermining about Plum1_N's situation.[15] He could have been moved to act by any number of factors—for instance he might have just have seen a motivational poster that told him to *just do it* (i.e. kill White) and he might have then killed White. The reason I don't find this responsibility-undermining is that Plum1_N is the sort of person who reasons egoistically. He often, independently from the neuroscientists, reasons this way, and it is perhaps only because of circumstantial luck that he wouldn't have reasoned this way if it were not for the neuroscientists. Thus it seems reasonable to suppose that he could have killed White independently of the neuroscientists *if* his circumstance had been slightly different. Because Case 1_N does not have same intuitive pull on the compatibilist as Case 1, a manipulation argument that is grounded by Case 1_N will thus *not* have a same dialectical power as a manipulation argument grounded by Case 1. Since I have argued that there is a metaphysically coherent interpretation of Case 1, Pereboom need not appeal to Case 1_N. He can thus retain his older, more powerful Four-Case Argument.

Compatibilists might, of course, fall back on their earlier soft-line replies, such as those defended by Baker (2006), Mele (2005), or Fischer (2004). But such replies will not succeed. Pereboom can easily change his Case 1 so that Plum1 is manipulated throughout his entire life, and thus satisfies the conditions proposed by these authors. This is sufficient to undercut these soft-line replies.[16] And, as I've argued, the reverse generalisation argument seems just as unhopeful.

As I've said, I'm a compatibilist. While I appreciate (and sometimes fear) the power of the Four-Case Argument, I think that compatibilists will eventually be able to rise to the challenge it presents them. But compatibilists will have to dig deeper to develop such a reply. While the reverse generalisation argument (on its own, at least) and the aforementioned soft-line replies are ineffective, there are perhaps other ways for compatibilists to defend themselves. They could try focusing on the details of Case 1, in a manner similar to Demetriou. While, as I've argued, it is not ultimately effective to argue Plum1 is not an agent, it is possible that there is

[15] Case 1_N is also similar to a case of Mele's (2009: 178). This case features Carl who has some medium strength desires to eat snacks, and who is then manipulated to act on those desires. Mele (2009: 178) contends that Carl *is* morally responsible for acting on those desires, despite the fact he has been manipulated. Note that Mele is agnostic between compatibilism and incompatibilism.

[16] Baker's reply does warrant further comment, as it is much more developed than either Mele's or Fischer's replies. She argues that Plum1 is not an agent because he lacks a first-person perspective (at least when he is being manipulated). However, her argument relies on the premise that (sporadic or non-sporadic) continuous manipulation cannot lead to an individual having a first-person perspective. Baker (2006: 320) supports this claim by arguing that it is not *physically* possible for brain-manipulation to instantiate the property of having a first-person perspective. But even if this is true, it certainly seems conceptually possible that brain-manipulation may instantiate such a property. The fact that manipulation arguments go beyond what is physically possible is not an adequate objection, since compatibilists are typically trying to provide *conceptually* necessary and sufficient conditions for being morally responsible. Given this, mere conceptual possibilities can affect what we take these conditions to be.

something else going in Case 1 which might explain away the intuition that Plum1 is not morally responsible. Alternatively, it might be that there are compatibilist conditions on moral responsibility that have not yet been proposed that might rule that Plum1 is *not* morally responsible. Or there might be a response available that combines both of these strategies. At this point, the onus is on the compatibilist to develop such a reply. Incompatibilists can rest easy until such a time.[17]

References

Baker, L. R. (2006). Moral responsibility without libertarianism. *Nous, 40*(2), 307–330.

Baron, R. (1997). The sweet smell of… helping: effects of pleasant ambient fragrance on prosocial behavior in shopping malls. *Personality and Social Psychology Bulletin, 23*, 498–503.

Beebee, H. (2002). Transfer of warrant, begging the question, and semantic externalism. *Philosophical Quarterly, 51*(204), 356–374.

Beebee, H., & Mele, A. R. (2002). Humean compatibilism. *Mind, 111*(442), 201–223.

Bennett, P., Hague, A., & Perkins, C. (1991). The use of Baker-Miller pink in police operational and university experimental situations in Britain. *International Journal of Biosocial and Medical Research, 13*, 118–127.

Demetriou, K. (2010). The soft-line solution to pereboom's four-case argument. *Australasian Journal of Philosophy, 88*(4), 595–617. **(see also Mickelson, Kristin)**.

Fischer, J. M. (2004). Responsibility and manipulation. *Journal of Ethics, 8*(2), 145–177.

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

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5–20.

Hume, D. (1739/1978). In P. Nidditch (ed.). *An enquiry concerning human understanding*. Oxford: Clarendon Press.

Kearns, S. (2012). Aborting the zygote argument. *Philosophical Studies, 160*(3), 379–389.

King, M. (2013). The Problem with manipulation. *Ethics, 124*(1), 65–83.

Lycan, W. (1987/1995) *Consciousness*. MIT Press: London.

Matheson, B. (2014). Compatibilism and personal identity. *Philosophical Studies, 170*(2), 317–334.

McKenna, M. (2008). A hard-line reply to pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research, 77*(1), 142–159.

Mele, A. (2005). A critique of pereboom's 'four-case argument' for incompatibilism'. *Analysis, 65*(285), 75–80.

Mele, Alfred. (2009). Moral responsibility and agents' histories. *Philosophical Studies, 142*(2), 161–181.

Mickelson, K. (2015). The zygote argument is invalid; now what? *Philosophical Studies, 172*(11), 2911–2929 **(see also Demetriou, Kristin)**.

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

Pereboom, D. (2005). Defending hard incompatibilism. *Midwest Studies in Philosophy, 29*(1), 228–247.

Pereboom, D. (2008). Defending hard incompatibilism again. In Nick Trakakis & Daniel Cohen (Eds.), *Essays on free will and moral responsibility*. Newcastle: Cambridge Scholars Publishing.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Profusek, P. J., & Rainey, D. W. (1987). Effects of baker-miller pink and red on state anxiety, grip strength, and motor precision. *Perceptual and motor skills, 65*(3), 941–942.

Schauss, A. (1985). The physiological effect of colour on the suppression of human aggression: Research on Baker-Miller pink. *International Journal of Biosocial Research, 2*(7), 55–64.

Schlosser, M. (2015). Manipulation and the zygote argument: another reply. *Journal of Ethics, 19*(1), 73–84.

[17] Thanks to Helen Beebee, Paul Russell, Catharine Abell, Derk Pereboom, and Natalie Ashton for feedback on earlier versions of this paper.

Shabo, S. (2010). Uncompromising source incompatibilism. *Philosophy and Phenomenological Research, 80*(2), 349–383.

Taylor, R. (1974). *Metaphysics*. Englewood Cliffs, NJ: Prentice-Hall.

Todd, P. (2013). Defending (a modified version of) the zygote argument. *Philosophical Studies, 164*(1), 189–203.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard: Harvard University Press.