The published version of this essay (which may differ slightly from the manuscript below) may be found here:

Kristin Demetriou (a.k.a. Kristin M. Mickelson) (2010). The Soft-Line Solution to Pereboom's Four-Case Argument, *Australasian Journal of Philosophy* 88.4: 595-617.

The Soft-Line Solution to

Pereboom’s Four-Case Argument

1. Introduction

For over a decade, compatibilists have struggled to respond to a powerful manipulation argument developed by Derk Pereboom: the notorious “Four-Case Argument”.[[1]](#footnote-1) Like other manipulation arguments, Pereboom’s is designed to refute compatibilism by pointing to a fundamental similarity between the effects of freedom- and responsibility-undermining manipulation and the effects of causal determinism. In the first stage of the argument, Pereboom attempts to show that an individual can satisfy a collection of the most famous compatibilist conditions for free will without satisfying the control requirements of moral responsibility. Using this strategy, Pereboom hopes to reveal that compatibilists have failed to capture even the minimal type of meaningful freedom—the type of freedom required for moral responsibility. While many other manipulation arguments stop there, Pereboom goes one step further, generating the remarkable power of the Four-Case Argument with the diagnosis that his manipulation victims lack the requisite amount of control for moral responsibility because their thoughts and behaviors are *causally determined* by their manipulators. Clearly, if this evaluation of the responsibility-undermining feature of the manipulation is correct, then the same responsibility-undermining feature is present in *every* action performed in a causally deterministic world. Thus, the Four-Case Argument not only threatens to discredit all known accounts of compatibilism, but also aspires to show that compatibilism is *in principle* a metaphysically untenable position.

In this paper I argue that it is Pereboom’s manipulation argument, and not compatibilism, that is untenable. I begin with a review of the Four-Case Argument, followed by a discussion of Michael McKenna’s valuable distinction between ‘hard-line’ and ‘soft-line’ replies to arguments of this kind. I quickly depart from McKenna’s treatment of the 4-CA, however, because his preferred hard-line reply fails to address many plausible, and arguably the most charitable, interpretations of the argument. More than one relevant interpretation of the argument is available, I claim, because there is an important ambiguity in the description of the causal relations found in the argument’s foundational case, Case 1. In an effort to resolve this ambiguity and, thereby, make a final evaluation of the 4-CA possible, I employ my endeavor to identify all of the metaphysically coherent resolutions of this ambiguity. For each interpretation I present, I argue that it falls under one of the two horns of a dilemma. The upshot of the dilemma, I contend, is that for every possible interpretation of the 4-CA, the compatibilist is able to provide either a compelling hard-line or soft-line response to it. Since there is no interpretation of the 4-CA which cannot be answered, I conclude that the 4-CA’s general attack on compatibilism fails.

In the next section, I consider a plausible alternative characterization of the dilemma which seems, at first blush, to breathe new life in the deflated 4-CA. In light of this recharacterization, it seems as though all of the hermeneutically viable interpretations of Case 1 support the 4-CA’s generalization strategy, meaning that the 4-CA can still be used to show that compatibilism is in principle untenable. In response, I diagnose the root problem with all of the viable interpretations of Case 1, showing that each of these interpretations involves a type of manipulation which undermines the victim’s agency. If my diagnosis is correct, it would mean that every viable interpretation suffers from the same basic defect and, so, would invite the same soft-line reply. When this collection of soft-line replies is taken as a whole, it becomes much more powerful than any one of its members—so powerful, in fact, that it provides the compatibilist with a *solution* to the 4-CA. Finally, I step back and present the foundations for a new taxonomy of manipulation arguments. I locate the Four-Case Argument in this taxonomy and conclude that any manipulation argument employing its type of manipulation is categorically defeated by the considerations I have offered.

2. The Design of the Four-Case Argument

To get the argument started, Pereboom collects five of the most popular “causal integrationist conditions” that have arisen out of the compatibilist camp. Summarizing each of Pereboom’s descriptions into slogan form, the five conditions are constancy of character, lack of constraint by irresistible desire, proper conformity of first-order and second-order desires, the capacity to regulate one’s behavior based upon a moderately reasons-responsive deliberation process, and the capacity to understand and regulate one’s behavior based on moral reasons. Pereboom labels these “integrationist conditions” because each is designed to capture a type of integration between an agent’s psychology and his actions necessary for an agent to have sufficient control to be a candidate for moral responsibility.[[2]](#footnote-2)

Notably, Pereboom emphasizes that the compatibilist causal integrationist conditions for freedom are not expected to be sufficient for moral responsibility entirely on their own. In other words, a compatibilist is not responsible for giving a *complete* analysis of moral responsibility. When a philosopher provides compatibilist conditions for moral responsibility, his main goal is to provide conditions that confirm the compatibility of determinism with the type of freedom or control required for moral responsibility, though there are also some further “implicitly understood (non-incompatibilist) conditions about agency, knowledge, and circumstance” that must be satisfied as well (Pereboom 2002: 111). As would be expected, Pereboom stipulates that the aforementioned set of background conditions for moral responsibility are satisfied in each of his four cases, in addition to the specific compatibilist conditions he is targeting.

In the first stage of the Four-Case Argument (hereafter, the “4-CA”), Pereboom offers two cases of manipulation that are designed to show that an agent can satisfy the compatibilist integrationist conditions and yet fail to be morally responsible for his behavior. His goal is to provide a case in which an individual is subjected to an intuitively freedom- and moral-responsibility-undermining form of manipulation but still satisfies the compatibilist integrationist conditions, which would establish that even the best and the brightest of the compatibilists have failed to provide sufficient conditions for the freedom required for moral responsibility. Having shown that the compatibilists have failed *so far,* Pereboom’s argument would indicate a looming threat for any future compatibilist account of freedom: no matter what further condition a compatibilist might concoct to complete the set of sufficiency conditions, a manipulation argument is waiting in the wings to undermine it.

The first putative counterexample features an individual, Plum, who is designed by neuroscientists so as to satisfy the compatibilist integrationist conditions and yet does not seem morally responsible for his actions:

Case 1. Professor Plum 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 is possible, given this 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 process exemplifies the various components of moderate reasons-responsiveness. He is receptive to the relevant pattern of reasons, and his reasoning process would have resulted in different choices in some situation in which the egoistic reasons were otherwise. At the same time, he is not exclusively rationally egoistic since he will typically regulate his behavior by moral reasons when the egoistic reasons are relatively weak—weaker than they are in the current situation (2002: 113).

In this case, all of the compatibilist integrationist conditions appear to be satisfied, but the features of Plum that satisfy the five conditions have been covertly manipulated into place. The manipulation is clearly quite severe: during the period of manipulation, the neuroscientists directly cause Plum’s every state—at least every state of his reasoning process—on a moment-to-moment basis.[[3]](#footnote-3) Due to the nature of the manipulation, the intuitive response to Case 1 from most compatibilists is that Plum is *not* morally responsible when he finally kills Ms. White.[[4]](#footnote-4) The best explanation for this intuition, Pereboom claims, is that Plum’s murderous act was beyond his control. More specifically, Pereboom argues that our assessment that Plum’s behavior is beyond his control is best explained by the fact that the behavior was *causally determined* by the neuroscientists. Indeed, no other compelling explanation seems readily available.

Worried that compatibilists might argue that Plum in Case 1 (hereafter, “Plum1”) is not morally responsible because of the moment-by-moment aspect of the neuroscientists’ control over his behavior, Pereboom adds a time lag to the control exerted by the neuroscientists on Plum to create Case 2:

Case 2. Plum is like an ordinary human being, except that he was created by neuroscientist, who, though 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 (2002: 113-14).

As in Case 1, the intuitive response is that Plum in Case 2 (“Plum2”) is not morally responsible for killing Ms. White because his murderous act was beyond his control, having been causally determined by the neuroscientists. Thus, despite the addition of the time lag, Pereboom’s original argument to the best explanation seems to hold. Having established that the time lag makes no difference between Case 1 and Case 2 in terms of moral responsibility, Pereboom states that Case 2 alone is a satisfactory counterexample to the sufficiency of the compatibilist integrationist conditions. Thus, if *either* Case 1 or Case 2 is successful, so is the first step of Pereboom’s argument.

In the second stage of the argument, Pereboom employs a generalization strategy, constructing a bridge case from his purported counterexample cases to the case of a normal human in a deterministic world. Given that my critique of the 4-CA focuses almost entirely on Case 1 and Case 2, I will forgo a detailed review of the final two cases. Suffice it to say, the bridge case, Case 3, is a near-normal situation in which overbearing parents impose rigorous training on young Plum. Pereboom expects that the intuitive response to Case 3 will be that Plum (“Plum3”) *is* morally responsible for murdering White, despite the rigor of his training. The trouble for compatibilists is that there seems to be no principled difference between the first two cases and Case 3 that could justify holding Plum3 responsible while denying of responsibility to Plum1 and Plum2.

Worse yet, if Pereboom is correct that the responsibility-undermining feature of Case 1 and Case 2 is the fact that the victim is unable to control his behavior because it is causally determinedby the neuroscientists, then it appears that the compatibilist will be forced to admit that another Plum, one embedded in a causally deterministic world, cannot be responsible for any of his actions either! To drive home this point, Pereboom concludes his generalization argument with the presentation of a fourth Plum (“Plum4”) who is embedded in a causally deterministic world. Though Plum4 satisfies the compatibilist integrationist conditions and is intuitively a free and responsible agent, the responsibility-undermining feature identified in Case 1 and Case 2 is present in Case 4, i.e. Plum4’s actions are causally determined. Without a principled way to distinguish Plum4 from the other Plums, our moral assessment of Plum4 must align with our assessments of the previous three Plums. The compatibilism-refuting conclusion now seems unavoidable: Plum4 is not morally responsible for killing White.

3. Discourse on the Dialectic

In a recent article, McKenna recommends a general strategy for compatibilists wishing to respond to manipulation arguments such as the 4-CA (McKenna 2008). He suggests that the compatibilist has two options: she can pursue either a “hard-line” or “soft-line” reply. Defenders of the hard-line start by seeking out an interpretation of the manipulation which, in keeping with the spirit of Pereboom’s stipulations, satisfies all of the conditions that the hard-liner considers necessary for free and responsible agency. Of course, as McKenna explains, once it is clear that the manipulation victim satisfies all of these conditions, the compatibilist can reasonably respond that Pereboom’s manipulation victims are free and morally responsible after all, and thereby undermine the 4-CA.[[5]](#footnote-5) By contrast, soft-liners start by accepting an interpretation of the manipulation which generates the key intuition that the manipulation is responsibility-undermining. Thus, soft-liners accept the challenge of showing how the manipulation victims *differ* from agents who are free and responsible. In order to meet this burden, it seems that the soft-liner must either (1) reveal that Pereboom’s manipulation victims fail to satisfy a condition which she has previously claimed to be necessary for free and/or responsible agency, or (2) provide a new (but not *ad hoc*) condition which establishes a principled, freedom- or responsibility-relevant difference between the manipulation victims and individuals embedded in a deterministic world. McKenna ultimately endorses the hard-line strategy, arguing that it is impossible for the compatibilist to develop a successful soft-line reply.

Like McKenna, I believe that a hard-line must be taken in responding to the 4-CA—but I do not think that the compatibilist can take an *exclusively* hard-line. While a detailed critique of McKenna’s view of the dialectic cannot be undertaken here, one of the basic mistakes underlying his conclusion is that he believes the 4-CA can be answered with a single hard-line reply. In fact, no single hard-line reply could be sufficient to answer the challenge of the 4-CA because there are multiple ways of interpreting the manipulation described in Case 1 of the argument and each requires its own response. As we shall see below, there are several interpretations which successfully neutralize the *prima facie* intuition that the manipulation victims are not morally responsible, making each a candidate for a distinct hard-line reply. However, even a collection of these hard-line replies would fail to provide an adequate response to the 4-CA because there are other interpretations of Case 1 which only serve to solidify the intuition that the manipulation robs its victim of moral responsibility. For each of these latter interpretations the challenge of the 4-CA remains: Can the compatibilist explain why the manipulation victims are not morally responsible without undermining her preferred version of compatibilism in the process? No hard-line reply can meet this challenge; only a soft-line reply will do.

4. The Causal Control Dilemma

Pereboom’s description of the manipulation in the foundational case of the 4-CA, Case 1, is obviously quite vague. Of course, many of the fine details that one might add to flesh out these manipulation scenarios are of little import to the overall argument and, rightly, Pereboom does not dwell on such minutiae. However, the success of the argument does depend on there being at least one coherent way to flesh out the metaphysical details of the responsibility-undermining manipulation, and it is not obvious that this can be done. The worry arises from an odd tension between some of the key details in Case 1: on the one hand, the neuroscientists are responsible for “directly producing (Plum1’s) every state”, but on the other hand we are told that “(Plum1) will…regulate his behavior”. The tension between these two stipulations is only increased when Pereboom explicitly states that the neuroscientists exercise *causal control* over Plum1’s actions which allows them to regulate Plum1’s behavior. After all, it should be uncontroversial that in order for Plum1 to “regulate” his own behavior, he too must exercise some minimal causal control over his actions. As it stands, then, Plum1 and the neuroscientists seem to be competing for causal control of Plum1’s states, leaving the exact nature of the manipulation far from clear. What is clear, however, is that the success of the 4-CA cannot be properly evaluated until this crucial ambiguity is resolved. This is because Case 1 will support the 4-CA only if there is a specificaccount of the relation between the causal contributions of the neuroscientists and those of Plum1 in producing Plum1’s act of murder which explains how it would be possible for Plum1’s causal contribution to be sufficient for him to causally regulate his own behavior despite the independent causal control exerted over him by the neuroscientists.

I believe that a dilemma looms for the proponent of the 4-CA who takes up the challenge of resolving this ambiguity and providing this specific account.[[6]](#footnote-6) If the proponent offers an account on which Plum1 could be said to ‘win’ this causal competition such that Plum1 exerts independent causal control over his own behavior, then the compatibilist can reasonably counter with a hard-line reply. That is, once it is clear that neuroscientists lack the causal power to interfere with Plum1’s causal control over his own states, the (so-called) manipulation would be so innocuous that there would be no remaining reason to think that Plum1 is not a candidate for moral responsibility. But suppose on the other hand that the proponent of the 4-CA adopts an interpretation on which Plum1 ‘loses’ the causal competition. Even though this sort of interpretation would generate the key intuition that Plum1 is not responsible, it would do so in virtue of Plum1’s violating the two compatibilist integrationist conditions requiring that he be able to self-regulate—all the required ingredients for a soft-line reply. In order for the 4-CA to threaten compatibilism, there must be an account of the manipulation that avoids both horns of this dilemma, what I call “The Causal Control Dilemma”, by somehow granting Plum1 the causal power to regulate his own behavior while yet generating the intuition that Plum1 is not morally responsible.

The proponent might try to avoid each horn by appeal to overdetermination: perhaps *both* Plum1 and the neuroscientists exert independent and equally efficacious causal power in bringing about Plum1’s each and every state. At this point, we could quickly get mired in a discussion about the viability of overdetermination, getting bogged down in the controversy over the metaphysical possibility of overdetermination in isolated instances, let alone the possibility of the pervasive overdetermination that would be required for Plum1 and the neuroscientists to overdetermine Plum1’s every state. Luckily, a journey into that treacherous territory is avoidable, given that Pereboom could not successfully appeal to overdetermination to explain the causal relation between Plum1 and the neuroscientists. Given the details of Case 1, Pereboom would have to be talking about perfect state-by-state overdetermination: Plum1 and the neuroscientists never diverge in purpose, with the result that Plum1 never has a non-overdetermined state during the manipulation. So, even if we imagine (for simplicity) that the overdetermination interpretation leaves open the physical possibility that Plum1 could have attempted to do something other than what the neuroscientists caused him to do, it just so happens that he never does make such an attempt; even in the absence of the causal contributions of the neuroscientists, Plum1 would have behaved no differently. Indeed, by its very definition, the overdetermination interpretation guarantees that Plum1’s causal contribution alone is *sufficient* to bring about all of the states leading up to the murder. So, even if Plum1’s states are overdetermined, it would seem reasonable to conclude that Plum1 has sufficient control to be morally responsible for that murder—a hard-line reply.

Perhaps causal interactionism provides a more promising escape-route from the dilemma posed above? On this strategy, the scientists and Plum1 are each causes of Plum1’s behavior in virtue of being alternating links on the same causal chain which brings about Plum1’s states. McKenna seems to endorse an interpretation of this kind in mounting his hard-line response to the 4-CA, suggesting that one might consider the neuroscientists to be “causal prosthetics”, transmitting causal messages between Plum1 and his environment and, presumably, between Plum1’s states as well. “On this model”, says McKenna, “while (the neuroscientists are) able to steer Plum in certain directions (like to kill Ms. White), often times, (the neuroscientists are) functioning merely as a sort of extra causal link in a chain. (The neuroscientists function) like a prosthetic, allowing Plum to deal with his world like any other agent” (McKenna 2008: 149-50).

Following McKenna’s lead, let us consider a case in which the neuroscientists are slavish causal prosthetics who faithfully convey causal signals between Plum1’s states, such that the neuroscientists cause precisely the same states in Plum1 as Plum1’s antecedent states would have caused by themselves in the absence of the neuroscientists. In effect, the neuroscientists employ their causal powers in the service of Plum1, so it seems that they once again lose the competition for causal control over Plum1 states—only Plum1 truly controls or regulates his behavior. Once the compatibilist is convinced that Plum1 exercises such control of his behavior, though, it is likely that her intuition that Plum1 is not morally responsible will dissolve. Indeed, McKenna worries that this interpretation makes it *so* obvious that Plum1 is a morally responsible agent that adopting it might be seen as reducing the 4-CA to a non-starter (McKenna 2008: 150, fn. 6). Viewing the neuroscientists as ‘faithful prosthetics’, then, apparently leads to another compelling hard-line reply to the 4-CA.

Equally problematic, however, is the scenario in which the neuroscientists interpose themselves between Plum1’s states but *fail* to act as perfectly faithful causal prosthetics, such that they cause Plum1 to behave *differently* than his prior states would have caused him to act. Admittedly, such an ‘*unfaithful* causal prosthetic interpretation’ is in the spirit of Case 1. It fits well with claims like, “The neuroscientists manipulate (Plum1) 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*” (Pereboom 2002: 113; italics added). This passage describes Plum1’s reasoning as being causally initiated by the neuroscientists, regardless of what would have followed naturally from Plum1’s prior states, so that Plum1 thinks and behaves any way the neuroscientists decide. Understanding the manipulation in this way would surely lead to the key intuition that Plum1 lacks moral responsibility for his actions, which means that it generates the intuition that the 4-CA depends upon. The problem with the unfaithful prosthetic approach, though, is that when the neuroscientists are unfaithful in conveying the causal signals between Plum1’s states, the neuroscientists once again win the competition for causal control of Plum1’s states. Since on this interpretation it is impossible for Plum1 to exercise causal control over his own behavior, it is not amenable to a hard-line reply. It does, however, suggest a soft-line reply based on the fact that Plum1 does not satisfy all of the compatibilist integrationist conditions. Thus, the 4-CA offers no genuine threat to compatibilism on either the faithful or unfaithful prosthetic interpretations.

Only one interpretation of the causal relations underlying the dual regulation of Plum1’s behavior seems left to discuss: one wherein the neuroscientists and Plum1 compose a jointly sufficient cause for each of Plum1’s states, i.e., one on which neither the scientists nor Plum1 alone is sufficient to bring about Plum1’s states, and only together are they able to bring about Plum1’s states.[[7]](#footnote-7) Unfortunately for Pereboom, a closer look reveals that the jointly sufficient cause interpretation gives rise to a similar dilemma to the one that undermined the causal prosthetic interpretation.

Pereboom tells us that Plum1 “is as much like an ordinary human being as is possible” (2002: 113), so it seems reasonable to assume that, if the neuroscientists had simply released Plum1 into the world upon his creation and performed no further manipulation on him then Plum1 would have been able to act like an ordinary human being. This, in turn, suggests that the causal contributions of Plum1’s states were designed to be sufficient to bring about his subsequent states. It seems, then, that the jointly sufficient cause interpretation could only work if the neuroscientists, as part of their manipulation of Plum1, *undermine* the causal sufficiency of Plum1’s states in some way. Now, one can imagine various stories about how the neuroscientists could do this, but the details will ultimately be of little import. When the neuroscientists assert their own causal powers to jointly cause Plum1’s behavior, they would have to do so in one of two ways: either the neuroscientists make Plum1 behave just as he would have behaved in their absence or they cause Plum1 to behave differently than he would have behaved in their absence. As discussed in the faithful prosthetic interpretation, when the neuroscientists use their causal powers to bring about exactly the same states in Plum1 as would have resulted in their absence, it seems reasonable for the compatibilist to believe that the neuroscientists’ influence does not undermine Plum1’s moral responsibility for his resulting actions—which is to say, the compatibilist can give a compelling hard-line reply. On the other hand, if the neuroscientists use their causal powers to make Plum1 act differently than he otherwise would have, the neuroscientists would thereby undermine Plum1’s moral responsibility. However, if the neuroscientists change Plum1’s behavior in the latter way, then the compatibilist can use the same argument used against the unfaithful prosthetic version of the interactionist interpretation discussed above. Namely, the compatibilist can offer the soft-line response that Plum1 is not morally responsible because the unfaithful changes to Plum1 made by the neuroscientists undermine his ability to regulate his own behavior—Plum1 would have done otherwise had only things been left up to him.

I hope that the gravity of the Causal Control Dilemma is now clear. On one hand, we have the interpretations of the manipulation on which Plum1 *wins* the competition for causal control of his states, retaining enough causal control that compatibilists would consider him morally responsible for his actions. On each of these interpretations, Case 1 fails to generate the intuition that is needed in order to run the 4-CA and, so, each invites a persuasive hard-line reply. On the other hand, we have the interpretations on which Plum1 *loses* the competition to the neuroscientists. Each of these latter interpretations leads to the intuitive response that the 4-CA depends on, making it necessary for the compatibilist to identify a responsibility-undermining feature of the manipulation. But we have seen that, in response to each case, the compatibilist is able to point to the same feature, which means that she can offer the same soft-line reply to each interpretation. In short, the manipulation victim fails to be morally responsible because he does not have the causal control required to self-regulate, and thus cannot satisfy the compatibilist integrationist conditions—conditions which a normal agent in a deterministic universe, like Plum4, could satisfy. Ultimately, it appears that there is no interpretation of the causal relations between Plum1 and the neuroscientists that can preserve both the stipulations and intuitions that the 4-CA depends on.

5. The Soft-Line Solution: Part One

But wait—Pereboom expects that the intuitive response to the story he tells in Case 1 will be that Plum1 is not morally responsible because Plum1’s behavior is causally determined and therefore beyond his control. Reflecting on these central features of the 4-CA, one might begin to wonder if the Causal Control Dilemma is really as devastating as it appears. First of all, each of the interpretations of Case 1 falling under the first horn of the dilemma, i.e. those in which Plum1 wins the causal competition, fail to generate the expected non-responsibility intuition. The success of the 4-CA straightforwardly depends on its ability to generate this intuition, so fleshing out Case 1 in accordance with any of these interpretations would reduce the 4-CA to a non-starter. Thus, even though each is a *metaphysically* *coherent* interpretation of Case 1, one might reasonably argue that each of these interpretations is so horribly uncharitable to the 4-CA that none can be considered a *hermeneutically* *viable* interpretation—especially in light of the fact that more friendly alternatives exist. Assuming this is right, and I believe it is, the Causal Control Dilemma should be seen first and foremost as separating the unviable interpretations from the viable ones. In light of this recharacterization, it becomes clear that each of the hard-line replies discussed above are directed at unviable interpretations of the 4-CA—so, properly speaking, they are not directed at the 4-CA at all—which means that they do not indicate any weakness in Pereboom’s argument. Ultimately, then, the success or failure of the 4-CA must be determined by the quality of the soft-line replies given to the viable interpretations of it.

Once we narrow our focus to the soft-line replies, though, the proponent of the 4-CA might insist that the content of these replies actually highlights the success of the most important aspect of the 4-CA: the generalization strategy. Upon review, the proponent might argue, the viable interpretations of Case 1 generate the intuition that Pereboom expects and, it seems, for precisely the reason that Pereboom identifies: Plum1 intuitively lacks the control required for moral responsibilitybecause his actions are *causally determined*. If this is right—and the compatibilist already seems to have agreed that it is—the 4-CA still leads to a conclusion that is devastating to compatibilism: Plum4, the normal agent in a deterministic world, lacks the control required to self-regulate and so cannot be morally responsible simply because his states are causally determined. Here, then, the original generalization strategy of the 4-CA is operating in full effect, apparently establishing that compatibilism is in principle an incoherent position. Now, in order to adopt this line of defense, one must sacrifice Pereboom’s claim that all of the compatibilist integrationist conditions are satisfied by Plum1, but this is hardly problematic. The proponent might easily argue that, in light of the success of its generalization strategy, the 4-CA not only shows that compatibilism is in principle false, but also that a determined agent cannot satisfy even the most anemic of the compatibilist causal integrationist conditions. It seems, then, that the 4-CA still points to an embarrassing flaw in (at least some) contemporary accounts of compatibilism while on its way to rule out all of them. This shows, one might conclude, that the 4-CA emerges virtually unscathed from the purportedly insoluble Causal Control Dilemma.

Fortunately for compatibilism, the compatibilist can block even this revitalized version of Pereboom’s generalization strategy. This is because there is an important difference between, on the one hand, cases in which one’s behavior is deterministically caused by such things as Pereboom’s brain-tweaking manipulators, and on the other hand cases in which one’s behavior is causally determined by one’s own prior states (as would be the case in a causally deterministic world). While others have made similar attempts to defend the existence of a morally relevant difference between these scenarios, supporters of the 4-CA have been less than impressed because, hitherto, the metaphysical underpinnings of this difference have not been adequately exposed. However, now that we know that the only interpretations of Case 1 which generate the intuition that Plum1 is not morally responsible are also those in which the neuroscientists win the competition for causal control over Plum1’s states, it is possible to expose the fundamental difference between the causal relations that obtain in the viable interpretations of Case 1 and those that obtain in Case 4. We have seen that when the neuroscientists win the competition for causal control of Plum1’s states, it is because the neuroscientists unilaterally initiate changes in Plum1’s states. With that in mind, consider the following diagrams illustrating the causal relations between the Plums’ bodily/brain states (B), the phenomenological mental states (M) associated with (B), and the manipulative neuroscientists (NS)[[8]](#footnote-8):

1a. Plum as Normal Human Person in a Causally Deterministic World (Case 4):

M1 M2 M3

Plum4

B1 B2 B3

2a. Plum as Causally Regulated by Neuroscientists (Case 1)[[9]](#footnote-9):

M1 M2 M3

Plum1

B1 B2 B3

*NS NS NS*

A deep difference between Plum1 and Plum4 is immediately apparent: Plum1 is not a causally integrated entity in the same way as Plum4.[[10]](#footnote-10)

Along a similar line, John Martin Fischer and Mark Ravizza have suggested that an individual like Plum1 might not be a “coherent self” and this explains Plum1’s lack of moral responsibility (Fischer 1998: 234-5, fn. 26). Although this response is intuitively compelling, it has been met with serious opposition. Pereboom claims that there is no reason to suppose that Plum1 is not a coherent self because “one might imagine that Plum’s mental states in Case 1 or Case 2 are qualitatively identical over time to those of a non-manipulated person” (Pereboom 2002: 121). Clearly, the above diagrams lend support to Pereboom’s response to Fischer and Ravizza, as they represent Plum1 and Plum4 as having the same qualitative experiences despite the differences in their circumstances. However, even if Plum1 and Plum4 have exactly similar physical and qualitative states, this does not ensure that Plum1 and Plum4 have the same status in terms of *agency*—a point that Fischer and Ravizza’s reply fails to drive home.[[11]](#footnote-11) With the above diagrams in hand, we can see now that even if we were to grant Pereboom’s point that a unified consciousness could arise from the manipulated brain in Case 1, and even if we were to grant that this entity had sufficient unity of conscious to be a coherent self, the compatibilist still has reason to reject that this ‘self’ is an agent. As displayed in Diagram 2a, Plum1’s physical and qualitative mental states are not causally efficacious in bringing about his subsequent physical and mental states; Plum1’s states are, rather, the end effects of the causal powers expressed by the neuroscientists. I take it to be uncontroversial that when the neuroscientists suppress the causal efficacy of Plum1’s states, taking the causal regulation of Plum1’s states into their own hands, that they thereby suppress his agency. With that in mind, I refer to this type of manipulation as “suppressive manipulation”. By contrast, a compatibilist would consider the causally integrated Plum4 depicted in Diagram 1a to be a paradigmatic agent.[[12]](#footnote-12) So, by attending to previously overlooked details, the compatibilist is finally in a position to identify a problem common to all of the viable interpretations of Case 1 that she has identified, a problem that does not generalize to Case 4.

Thus, it would appear that there is a significant difference between the effects of one’s being causally determined by suppressive manipulation and the effects of being an inhabitant of a causally deterministic world. What is more, since the neuroscientists undermine Plum1’s ability to self-regulate by disrupting his agency, it is plain that the compatibilist need not appeal to any of the controversial details of the causal integrationist conditions in order to give a decisive soft-line reply to every viable interpretation of the 4-CA. Since *any* compatibilist can endorse this series of soft-line replies, it seems that there is now a soft-line *solution* to Pereboom’s challenge: the principled difference between Case 1 and Case 4 is that Plum4 is a fully integrated agent but Plum1 is not (and could not be so long as the suppressive manipulation continues).[[13]](#footnote-13)

6. The Soft-Line Solution: Part Two

At this point, the reflective reader might notice that there is something suspect about the *prima facie* intuitions that I attribute to the compatibilist in the previous sections. Now that we have established that two individuals with exactly the same bodily and mental states can differ with respect to agency, it is no longer obvious that Plum1 is morally responsible just because the neuroscientists faithfully bring about the states in him that would have obtained in their absence. This means that the compatibilist will not be in a position to render a final judgment about Plum1’s moral responsibility in the scenarios where the neuroscientists faithfully produce Plum1’s states until she knows what accounts for the fact thatthe neuroscientists are faithful. In other words, the same details which were used to illuminate the problem shared by the “unfaithful” versions of the prosthetic interpretation and the jointly sufficient cause interpretation also indicate that the compatibilist should revisit their “faithful” counterparts. As we shall see, once these details are revealed, the compatibilist will have to reject her *prima facie* intuitive responses to these interpretations of Case 1.

Starting with the faithful prosthetic interpretation, recall that McKenna describes the neuroscientists as “functioning merely as a sort of extra causal link in a chain”, along with the use of the term ‘causal prosthetic’. This description gave McKenna’s interpretation the appearance of being an instance of the interactionist interpretation that we were looking for above. That is, it prompted us to imagine a causal chain in which Plum1’s states retain their causal efficacy. On this chain, the neuroscientists cause Plum1’s behavior insofar as they provide the proximate cause of Plum1’s thoughts and actions, but Plum1’s states are the proximate cause of the neuroscientists’ pressing the buttons they do, allowing for the judgment that Plum1’s states are causally responsible for his subsequent states. Viewing the case this way, which seems to be in the spirit of McKenna’s proposal, it appears obvious that Plum1 could be a morally responsible agent. Indeed, a diagram depicting this causal story would be relevantly similar to Diagram 1 above, showcasing Plum1 as a strange, but causally integrated agent.[[14]](#footnote-14) Thus, a *genuine* interactionist interpretation of the manipulation would generate an intuitive response that would make plausible the hard-line reply McKenna that offers. Unfortunately for McKenna, though, since this version of his interpretation fails to generate the crucial non-responsibility intuition, the import of his hard-line reply is arguably undercut by the fact that it responds to an unviable interpretation of Case 1.[[15]](#footnote-15)

However, we can also take McKenna’s description of his prosthetic interpretation at face value. When we do so, there are aspects of his description which make it incompatible with a straight-forward interactionist reading. Recall that McKenna allows the neuroscientists the flexibility to “steer” Plum1 as they see fit, when they see fit. So, while they might faithfully *choose* to cause precisely the same states in Plum1 that would have obtained in their absence, the neuroscientists might just as easily choose to initiate changes in Plum1 that would not have occurred in their absence. To see why it is problematic that the neuroscientists are able to choose which states they cause in Plum1, consider a period during which the neuroscientists fail to act as perfectly faithful causal prosthetics, such that they cause Plum1 to behave *differently* than his prior states would have caused him to act. In such a case, Plum1’s states are causally initiated by the neuroscientists, so that Plum1 thinks and behaves *any* way that the neuroscientists happen to decide. As shown in Diagram 2a, this sort of manipulation undercuts Plum1’s moral responsibility for his actions by undermining the causal integration required for Plum1 to be an agent. However, it should now be evident that *even if the neuroscientists happen to be perfectly faithful*, causing only those states in Plum1 that would have been caused naturally in their absence, Plum1 would still lack the causal integration required for agency! So long as the neuroscientists serve as the independent proximate causes of Plum1’s states, such that their button-pressings are expressions of their own desires rather than the effect of the causal powers exerted over them by Plum1’s prior states, it is Diagram 2a which accurately depicts the neuroscientists causal relation to Plum1.[[16]](#footnote-16) Assuming that the compatibilist should renounce their under-informed *prima facie* intuition to the interpretation McKenna describes rather than accept the absurd alternative that Diagram 2a depicts an agent, it seems clear that Plum1 is not morally responsible on a literal interpretation of McKenna’s prosthetic story because it fails to present Plum1 as an agent. Thus, on a strict reading, McKenna fails to achieve both his goal to present Plum1 as an agent and his goal to present an interpretation which could be used to support a hard-line reply to the 4-CA. Now, this should not overshadow the fact that the strict reading supports a viable interpretation of Case 1. However, because it is clearly one in which the neuroscientists subject to Plum1 to suppressive manipulation, the compatibilist can appeal to the same compelling soft-line reply she gave to the other viable interpretations of the argument.

By parity of reasoning, a soft-line reply is also fitting in the case where the neuroscientists faithfully offer their independent causal input to jointly cause Plum1’s states. Once again, the fact that they happen to use their causal powers in a faithful way does not create the causal integration required for Plum1 to be an agent; this interpretation, too, represents the neuroscientists as subjecting Plum1 to agent-undermining, suppressive manipulation. There seems little option but to admit that our *prima facie* intuition was misleading in this case, given that an individual cannot be morally responsible unless he is as an agent. As a result, the compatibilist must abandon the hard-line reply here as well, opting instead for the response that Plum1 is not morally responsible in this case because he is not an agent. So, once again, the compatibilist can adopt the same soft-line reply given to the other viable interpretations of the 4-CA.

Upon review, then, even after the compatibilist addresses the need to reject some of her *prima facie* judgments about Plum1’s moral responsibility, the Soft-line Solution to the 4-CA remains as strong as ever. The only difference is that the Soft-line Solution is now constituted by four instances of the same soft-line reply rather than two. Thus, after a grueling search for an interpretation of the 4-CA on which it poses a threat to compatibilism, we can finally conclude that there is none to be found.

7. The 3-Case Argument

The reader might wonder why the bulk of this paper is devoted to Case 1 given that Pereboom clearly states that Case 1 is a disposable part of his argument. The reasons are simple: Case 1 is easier to work with and all of the problems in Case 1 are inherited by Case 2. So, the arguments offered above against the 4-CA are equally successful against the remaining “3-Case Argument” (3-CA) which is based upon Case 2.

Case 2 is more difficult to understand than the first case for it includes an additional feature, the so-called “programming”, which is woefully under-described. Still, based on Pereboom’s commentary on Case 2, it is clear that Case 1 provides the guidelines for interpreting Case 2: Case 2 simply *is* Case 1 with a time lag. Pereboom adds the time lag to Case 2 precisely because he predicts that a compatibilist might come along who has worries about Plum1’s agency. Pereboom incorporates the time lag to appease such compatibilists, but is adamant that the time lag does not change anything of consequence. Reflecting on the small addition, Pereboom asks: “could a time lag between the manipulators’ activity and the production of the relevant states in the agent plausibly make a difference as to whether the agent is morally responsible? (…) By my intuitions, such a time lag, all by itself, could make no difference as to whether an agent is morally responsible” (Pereboom 2002: 113). I could not agree more with Pereboom on this point, but of course therein lays the problem.

Merely adding a time lag between the neuroscientists’ actions and the murder, such that “all the manipulating activity occurred during one time interval and, after an appropriate time lag, the relevant states were produced in the agent” (Pereboom 2002: 113) does not, all on its own, produce a morally relevant difference between Plum1 and Plum2. Presumably, then, since a difference in agency would be a morally relevant difference, it must be that Plum1 and Plum2 have the same status with respect to agency despite the presence of the time lag. But, given our earlier conclusion that Plum1 is not morally responsible due to his *lack* of agency, Pereboom’s own reasoning suggests that Plum2 should fare no better. Indeed, as the following diagram of Case 2 represents, the problematic aspect of Case 1 which I emphasized above, the *state-by-state* control that undermines Plum1’s agency, is still present in Case 2:

2b. Plum as Causally Regulated by Neuroscientists’ Program (Case 2):

M1 M2 M3

Plum2

B1 B2 B3

*P P P*

NS

So, while in Case 1 there was a tension between the neuroscientists and Plum1, in Case 2 there is an exactly similar tension between Plum2 and the neuroscientists’ programming. In Case 2, the program must regulate Plum2’s behavior, state by state and moment by moment, throughout his life, just as the neuroscientists directly regulate the behavior of Plum1. Thus, we can see that the suppressive manipulation which prevented the victim from being a candidate for moral responsibility in Case 1 is also present in Case 2.

I suspect that those who reject my interpretation of the causal relations in Case 2 will accuse me of misunderstanding the nature of the programming that the neuroscientists have implanted in Plum2 to do their dirty work. However, while there may seem to be ample room for debate about the nature of the programming, I believe the constraints on interpreting the programming are more limiting than it may first appear. Once one has Case 1 (the acknowledged template for Case 2) clearly in mind, it seems clear that the programming given to Plum2 must be *additional* to the basic programming that must have been present in Plum1. Although we have seen that Pereboom’s stipulation that Plum1 is an agent cannot be upheld because the neuroscientists undermine Plum1’s capacity for agency through their suppressive tweaking, it still seems reasonable to imagine that Plum1 is designed in such a way that, at the very least, Plum1 *would have been an agent* had the neuroscientists simply left him alone after his creation. From the fact that the neuroscientists need to send constant radio signals in order to carry out their manipulation of Plum1, it seems clear that the basic programming that was required to make Plum1 a functioning instant agent was not sufficient to provide the neuroscientists with the control over Plum1 that they desired. This suggests that the programming discussed in Case 2 must do something more than the basic programming given to Plum1; it must be something which allows the neuroscientists to get the thoughts and behaviors that they want from Plum2 on a moment-to-moment and state-by-state basis without the hassle of constant moment-by-moment monitoring and tweaking. In other words, the programming in Case 2 is designed to carry out the same type of suppressive manipulation that was achieved by the neuroscientists in Case 1, a type of manipulation that (assuming my arguments have been successful) always undermines agency. Ultimately, Pereboom’s description of Case 2, informed by our understanding of Case 1, seems to leave little room for doubt: Plum2, like Plum1, is the victim of suppressive manipulation and is therefore not an agent. Consequently, the 3-CA offers no threat to compatibilism.

8. The *New* 3-CA and Beyond

Of course, were a proponent of the 3-CA to jettison Case 1 from consideration, then she would immediately be free to interpret Case 2 however she likes. So, even if I have provided successful criticisms of the available interpretations of Case 1 and Case 2 on a strict reading of Pereboom’s argument, there may be some alternative interpretation of Case 2 that deserves attention because it can avoid all of the foregoing criticisms of the 3-CA.

On the most (perhaps only) plausible reinterpretation of Case 2, one could take the programming that the neuroscientists give to Plum2 to be nothing additional to the minimal amount of programming that would be required for an instant agent to function just like a normal human in a deterministic world.[[17]](#footnote-17) That is, one might argue that Plum2 is the nonhistorical duplicate of Plum4, i.e. a normal human in a deterministic world—not only in terms of mental and physical states as discussed earlier, but in terms of agency as well. Assuming this interpretation of the programming, there would be no grounds to conclude that Plum2 is any less an agent than Plum4, nor would there seem to be any obvious reason to think that Plum2 could not satisfy the compatibilist integrationist conditions. Presumably, the proponent of this version of the3-CA—let us call it “The New 3-CA”—would then point out that most people do not believe that an individual created in this strange way could be morally responsible for his actions. Thus, such an interpretation seems to provide all of the necessary components for avoiding the criticisms hitherto presented in this paper while satisfying the needs of a successful manipulation argument.

I openly admit that the New 3-CA avoids my criticisms of the original 3-CA and 4-CA. However, I contend that the New 3-CA is not merely a case of making the old argument better; rather, it is a new and better argument. Currently, compatibilists have no principled or systematic way of individuating manipulation arguments, which means that they have no good way of distinguishing between a shift from one *version* of an argument to another and a shift from one *argument* to another. I believe that the best and most natural way of individuating manipulation arguments is on the basis of the specific type of manipulation they employ, as it seems that all arguments involving the same type of manipulation will be subject to the same criticisms. Now, the type of suppressive manipulation described in the original interpretations of Case 1 and Case 2 is distinctive precisely because the manipulators continue to causally infect the states of their victim, moment by moment and/or state by state. In fact, the type of manipulation Plum1 and Plum2 are subjected to—what I would call suppressive “Create & Tweak Manipulation” because of the on-going involvement by the manipulators—is not employed in any other of the well-known manipulation arguments.

On the other hand, there are already a large number of manipulation arguments which involve the creation of a so-called “instant agent” who is immediately released into the world after his creation and is not tinkered with any further by his creators—what I refer to as “Create & Release Manipulation”. This type of manipulation is found, for example, in familiar cases developed by Alfred Mele (“Fred”), David Zimmerman (“Sean Young”), and Michael McKenna (“Suzie Instant”).[[18]](#footnote-18) In fact, part of the reason that compatibilists have been so troubled by the 4-CA is that they have been unable to extend their criticisms of the commonplace Create & Release manipulation arguments—where I happen to think hard-line responses alone are often adequate—to the Create & Tweak manipulation employed in the 4-CA. That is, it seems that the resiliency of the original 4-CA has come by way of the novel type of manipulation it employs, for it requires an equally novel response. Thus, once the defender of the 4-CA *abandons* the unique, suppressive Create & Tweak manipulation employed in the original 4-CA in favor of the non-suppressive Create & Release manipulation described in the New 3-Case Argument, he in fact concedes that the original 4-CA must be discarded—and, in that case, the New 3-Case Argument can be immediately relegated to the already burgeoning collection of manipulation arguments which employ Create & Release Manipulation.

Now that we have the beginning of a taxonomy of manipulation arguments, the compatibilist can confidently eliminate the entire category of suppressive manipulation arguments (such as the original 3-CA and 4-CA) from future debate. The task remaining for the compatibilists is to complete the taxonomy so that specific responses to each type of argument can be developed. Hopefully, by following this strategy the compatibilists will be able to force their opponents to retreat to an ever-smaller collection of arguments until no viable options remain.

9. Conclusion

Taken together, I believe my arguments not only show that the Four-Case Argument fails to reveal any inadequacy in contemporary compatibilism, but they also show that no future manipulation argument that employs suppressive manipulation will have any hope of succeeding. Not to be overlooked is the fact that my critique of the Four-Case Argument does not depend on the acceptance of any particular free-will machinery or any special theory of moral responsibility. In other words, the solution offered in this paper is not a mere circling-of-the-wagons defense of a particular version of compatibilism; it is designed to be a thorough-going refutation of Pereboom’s argument. I grant, however, that even if my arguments are sound, there is still a great deal of work remaining for the compatibilists. Even if suppressive Create & Tweak manipulation arguments have been categorically defeated, there are many other types of manipulation arguments which are not subject to the same criticisms, and the majority of these still go without a satisfying response.[[19]](#endnote-1)

References

Fischer, John Martin 2004. Responsibility and Manipulation, The Journal of Ethics 8/2: 145–77.

Fischer, John Martin and Mark Ravizza 1998. Responsibility and Control: A Theory of Moral Responsibility,

Cambridge: Cambridge University Press.

Frankfurt, Harry 1969. Alternate Possibilities and Moral Responsibility, The Journal of Philosophy 66/23:

829–39.

Haji, Ishtiyaque and Stefaan Cuypers 2006. Hard- and Soft-Line Responses to Pereboom’s Four-Case

Manipulation Argument, Acta Analytica 21/4: 19–35.

Kim, Jaegwon 1993. The Nonreductivist’s Troubles with Mental Causation, in Supervenience and Mind,

Cambridge: Cambridge University Press: 336–57.

Kim, Jaegwon 1998. Mind in a Physical World, Cambridge, MA: The MIT Press.

Kim, Jaegwon 2005. Physicalism, Or Something Near Enough, Princeton: Princeton University Press.

McKenna, Michael 2004. Responsibility and Globally Manipulated Agents, Philosophical Topics 32: 169–92.

McKenna, Michael 2008. A Hard-Line Reply to Pereboom’s Four-Case Argument, Philosophy and

Phenomenological Research 77/1: 142–59.

Mele, Alfred 1995. Autonomous Agents, New York: Oxford University Press.

Mele, Alfred 2006. Free Will and Luck, New York: Oxford University Press.

Pereboom, Derk 1995. Determinism al Dente, Nouˆs, 29/1: 21–45.

Pereboom, Derk 2001. Living Without Free Will, Cambridge: Cambridge University Press.

Pereboom, Derk 2002. Living Without Free Will: The Case For Hard Incompatibilism, in The Oxford

Handbook of Free Will, ed. Robert Kane, New York: Oxford University Press: 477–88.

Zimmerman, David 1999. Born Yesterday: Personal Autonomy for Agents without a Past, Midwest Studies

in Philosophy 23: 236–66.

1. The original version of the argument is presented in Pereboom’s “Determinism al Dente” in *Noûs*, 1995. However, it is the now standard version developed in his 2002 book, *Living Without Free Will*, that will be addressed in this paper. [↑](#footnote-ref-1)
2. For a more detailed summary of the origins and details of these five conditions, see Derk Pereboom, *Living Without Free Will*,pp. 100-10. In brief, constancy of character and lack of constraint by irresistible desire are traditional compatibilist requirements from Hume, the latter also associated with A.J. Ayer. The third condition, requiring the conformity of higher and lower desires, is taken from Harry Frankfurt’s famous hierarchical account of the freedom required for moral responsibility. The fourth condition, reasons-responsiveness, is based primarily on the account of compatibilist control offered by John Martin Fischer and Mark Ravizza, while the specific requirement for responsiveness and regulation by moral reasons is from Jay Wallace. [↑](#footnote-ref-2)
3. As mentioned above, Case 1 is open to a wide variety of interpretations, which will be the focus of the next section of this paper. However, there are notable ways in which Case 1 is *not* ambiguous. For instance, Pereboom’s story clearly states that the manipulation is carried out “moment by moment”, i.e. over some extended period of time, which effectively rules out the possibility of viewing the manipulation as occurring all in one instant. Next, Pereboom makes it adequately clear that the states constituting Plum1’s reasoning process are affected by the manipulation rather than, say, just the reasons and desires upon which he reasons. This is confirmed by Case 2, where Pereboom says that the programming—which is offered as a perfect substitute for the neuroscientists—causally determines the way that Plum2 will “*weigh* reasons for action” (italics added). Finally, it must be that the neuroscientists use the radio signals to “directly” cause Plum1’s brain states, since sending the radio signals to anything but his brain would be a quite *indirect* way to tamper with Plum1’s state of mind. So, any manipulation story which does not involve the direct causal determination of the victim’s brain states, specifically those constituting his process of reasoning, would stand in conflict with the manipulation case that Pereboom describes and, therefore, would fail to be a hermeneutically viable interpretation of Case 1. (As a general point, I believe that it is extremely important to avoid taking unwarranted liberties in interpreting manipulation cases, which means that one should carefully distinguish viable interpretations of a given case from nearby manipulation scenarios which may also be quite interesting. For further discussion of this point, see Section VIII of this paper, “The *New* 3-CA and Beyond”.) [↑](#footnote-ref-3)
4. John Martin Fischer is a high profile (semi-)compatibilist who rejects this intuition. In a direct response to Case 1, Fischer states that “Professor Plum, it seems to me, is not blameworthy, even though he is morally responsible” (Fischer, “Responsibility and Manipulation”, *The Journal of Ethics 8*,p. 158). I believe that the arguments provided in this paper provide a way for proponents of Fischer’s account of freedom to respond to the 4-CA without appealing to the controversial claim that Plum1 is morally responsible for the murder of Ms. White. [↑](#footnote-ref-4)
5. For a categorization of some of the most famous attempts to respond to the 4-CA along the soft/hard divide, see “Hard- and Soft-Line Responses to Pereboom’s Four-Case Manipulation Argument” by Ishtiyaque Haji and Stefaan Cuypers, in *Acta Analytica* 21, 2006, pp. 19-35. [↑](#footnote-ref-5)
6. The general argument articulated in this section was inspired by Jaegwon Kim’s “Explanatory/Causal Exclusion Problem”. See, for instance, Jaegwon Kim, *Mind in a Physical World* (1998), and, “The Nonreductivist’s Troubles with Mental Causation”, *Supervenience and Mind* (1994). [↑](#footnote-ref-6)
7. The reader may note that a supervenience relation has not been discussed. This is because supervenience is a non-starter in this context. Given the details of Pereboom’s story, it seems that if Plum1’s states and causal powers were supervenient on those of the neuroscientists, then Plum1’s states and causal powers would simply reduce to those of the neuroscientists and Plum1 would clearly lack the independence to be a morally responsible agent. In order to a defend a *non-reductive* account of Plum1’s supervenient causal powers, a.k.a. “strong emergentism”, one would have to provide a positive metaphysical story of how it is possible for such new and independent causal powers to emerge from and then causally influence the subvenient base. At best, the proponent of this view would have to solve Kim’s Exclusion Problem before appealing to this type of relation to save the 4-CA. [↑](#footnote-ref-7)
8. This style of diagram is often used in philosophy of mind to represent different visions of mental causation. I believe that my arguments are effective regardless of one’s preferred theory of mental causation, so I leave it to the reader to fill in and deal with the unrelated challenges resulting from his/her views on mental causation. [↑](#footnote-ref-8)
9. One might wonder how information about Plum1’s states is transmitted to the neuroscientists in this scenario, given that such transmission presumably would be causal and no causal route running from Plum1 to the neuroscientists is represented. To clarify, the lack of a causal arrow here simply reflects the lack of a *direct* causal relation between Plum1’s bodily state and the subsequent button-pressing by the neuroscientists. That is, I do not want to deny the presence of a causal chain which could account for the neuroscientists knowledge of Plum1’s every state. What this diagram is designed to show that Plum1’s states are not the *proximate causes* of any of neuroscientists’ button-pressings (while, on the other hand, the proximate cause of any one of Plum4’s states is his own prior state). That is, 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 the neuroscientists are free to decide, for reasons all their own, which state to cause in Plum1 at any given moment of the manipulation. [↑](#footnote-ref-9)
10. This diagram will be useful even if one wishes to argue that Plum1 might have causal integration between some of his states even though his reasoning process and behavior is different than it would have been due to the causal input of the neuroscientists. In such cases, Diagram 2a could 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. (Of course, it is now highly suspect to call the causally disjointed series of states at issue a “reasoning process”.) This narrowing of scope does not affect my argument, of course, for the states constituting one’s reasoning process are the most central to one’s agency (at least the robust sort required for moral responsibility), and so a failure of causal integration among these states alone would be sufficient to undermine Plum1’s agency. [↑](#footnote-ref-10)
11. The larger problem with Fischer’s strategy, of course, is that it offers no clear interpretation of the manipulation cases, it provides little argument in favor of the general interpretation it assumes, and, so, leaves open the possibility of a more charitable reading of the cases which could side-step his criticisms. By contrast, the strategy in this paper is to leave *no* possible interpretation without a definitive response. [↑](#footnote-ref-11)
12. More precisely, Plum4 provides an uncontroversial *base* for being a paradigmatic agent, insofar as his states are causally efficacious in bringing about his subsequent states. The remaining details of how his mental states are interrelated and related to his physical body must yet be filled in some appropriate way. However, the crucial point is that while such details feasibly can be filled in for Plum4; by contrast, neither the physical nor the mental states of Plum1 are causally efficacious in regulating Plum1’s behavior. [↑](#footnote-ref-12)
13. *Even for most libertarians*, an individual would not qualify as an agent if he has no causal control over his choices or bodily movements. So, as long as the incompatibilist agrees that some causal contribution to one’s subsequent states is required for agency (at least among the states constituting one’s process of reasoning, as mentioned above in footnote 10), the incompatibilist should agree that the defense offered here is not driven by particularly compatibilist commitment—whether deterministic or indeterministic, the causal connections between Plum1’s states are suppressed by the neuroscientists. This fact may be of interest to libertarians who endorse event-causal indeterminism, for Pereboom has claimed that such libertarians and compatibilists are in the same sinking boat when it comes to answering the challenge of manipulation arguments like the 4-CA (see, for example, “Living Without Free Will: The Case For Hard Incompatibilism” (in *The Oxford Handbook of Free Will*, New York: Oxford University Press, 2002), p. 478. [↑](#footnote-ref-13)
14. In order to construct this diagram, one would simply have to (1) redraw Diagram 1a, (2) add an ‘NS’ between B1 and B2 and again between B2 and B3, and (3) insert an arrow of causation between the latter symbols to generate a new version of the causal chain represented in Diagram 1a. [↑](#footnote-ref-14)
15. To be clear, though, the critique of the 4-CA being developed in this section does not depend on whether this interpretation is in fact unviable—although I believe it is. For anyone who thinks that it *is* a hermeneutically viable interpretation, the fact that a hard-line reply to the 4-CA would be forthcoming is, all on its own, sufficient to preclude any hope of utilizing this interpretation to save the 4-CA. [↑](#footnote-ref-15)
16. Notably, once we abandon the idea that Plum1’s states *cause* the neuroscientists to press the buttons which then *cause* his subsequent states, then it no longer seems appropriate to say that Plum1’s states and the actions of the neuroscientists are alternating links on the same causal chain. Of course, if there is no causal chain, then the neuroscientists cannot simply be a strange but agent-preserving link in this chain. At this point, the neuroscientists no longer seem to behaving like any sort of prosthetic at all. [↑](#footnote-ref-16)
17. While the term ‘instant agent’ is typically used to refer to cases of *ex nihilo* creation, like Swampman, I’m assuming that Frankenstein-like creations like Plum1 and Plum2 also fall uncontroversially under this heading when they are said to awaken into life with all they need (physically, epistemically, and metaphysically) to be agents. [↑](#footnote-ref-17)
18. See David Zimmerman’s discussion of “Sean Young” in “Born Yesterday: Personal Autonomy for Agents without a Past” (1999); Alfred Mele’s discussion of “Fred” in *Autonomous Agents* (1995),and again in *Free Will and Luck* (2006);and Michael McKenna’s discussion of “Suzie Instant” in “Moral Responsibility & Globally Manipulated Agents” (2006). [↑](#footnote-ref-18)
19. I would like to thank Michael McKenna and Robert Hanna for their kind and insightful advice on multiple drafts of this paper. I would also like to thank the two anonymous referees for this paper at the *Australasian Journal of Philosophy*, whose excellent comments have resulted in great improvements. [↑](#endnote-ref-1)