Resisting the Manipulation Argument:
A Hard-liner Takes it on the Chin[1]

Manipulation arguments for incompatibilism have become all the rage as of late in debates about free will and moral responsibility.[2] This is for good reason. As I shall explain, given certain assumptions about the recent history and the present state of the debate, the dispute between compatibilists and incompatibilists looks as if it reduces to the controversy over manipulation arguments. Perhaps the most compelling version of an argument of this type is the one advanced by Derk Pereboom (1995, 2001), his well known four-case argument for incompatibilism. I am alone in criticizing Pereboom's argument by defending what I have termed a hard-line compatibilist reply (2004, 2008). Others have instead opted for a soft-line reply (e.g., Fischer, 2004; Haji, 1998, 2009; and Mele, 2005).

In this paper, I shall first situate the significance of manipulation arguments in relation to historical developments spanning roughly the last half century. Next I'll set out the basic structure of manipulation arguments in general, and consider compatibilists' options for resisting them. In doing so, I shall make clear why it is that, in my estimation, a soft-line reply to an argument such as Pereboom's is ill-advised. I'll then turn to Pereboom's particular formulation of a manipulation argument, set out my hard-line reply, and size up his intriguing (2008) response. Finally, I'll counter in a manner that, I hope, will bring to light a new set of questions as regards both manipulation arguments, as well as the debates over free will and moral responsibility more generally. My intention is to do more than merely reply to a critic. I mean to break new ground.

1. Terminological Preliminaries

It will be useful to start with a bit of terminology. As I understand it, *free will* is the unique ability of persons to control their conduct in the strongest manner necessary for moral responsibility.[3] A *free act* just is an act that issue from this ability. For present purposes, I shall follow Pereboom in understanding *moral responsibility* in terms of basic desert-entailing moral blameworthiness and praiseworthiness. According to Pereboom, an agent is *blameworthy* for an act, x, if she x-s freely, knowing that x-ing is morally wrong (2001: xx). The fact that an agent

---

[2] See, for example, Kane (1996), Pereboom (2001), and Mele (2006). For earlier formulations, see Wisdom (1934: 116); Taylor (1963: 45-6); and Locke (1975).

[3] Some, such as van Inwagen (2008), would strenuously disagree with this formulation, preferring instead to define free will in terms of the ability to do otherwise. But others, such as Mele (2006) and Pereboom (2001), operate as I do, focusing on the freedom that pertains to moral responsibility. Nothing rides on this dispute, however, in what follows. Everything crucial can be expressed exclusively in terms of moral responsibility and the freedom pertaining to it.

satisfies these conditions entails that she deserves blame for so acting. The *desert* at issue is basic in the sense that an agent deserves blame just because of her acting as she did, and not because of any further considerations, such as those that might flow from a consequentialist or contractualist rationale.[4] (Similar remarks apply for praiseworthiness.) *Determinism* can be understood as the metaphysical thesis that a complete statement of the (nonrelational) facts at any time in the remote past in conjunction with a complete statement of the laws of nature entails every fact about what happens at every time thereafter. *Compatibilism*, as I use the term, is the thesis that both free will and moral responsibility are compatible with determinism. If compatibilism is true, it is a metaphysical possibility that an agent perform a free act at a world in which determinism is true, and that she is morally responsible for doing so. *Incompatibilism* is the denial of compatibilism.

2. One Perspective on the Recent History of the Debate
   Incompatibilists have developed two different tactics in arguing against the compatibilist. One fixes upon a model of freedom cast in terms of different options open to a free agent whereby she is able to do otherwise. Another fixes upon a model of freedom cast in terms of the agent as the origin or initiator of her own conduct. These might be characterized, respectively, as *leeway* and *source* models of freedom. Here are two basic arguments that showcase the different strategies incompatibilists have adopted in their assault on compatibilism:
    Basic Leeway Incompatibilist (BLI) Argument:
      1. Free will requires the ability to do otherwise.
      2. If determinism is true, no one is able to do otherwise.
      3. Therefore, free will is incompatible with determinism.
    Basic Source Incompatibilist (BSI) Argument:
      1. Free will requires the ability to be the appropriate source of one's actions.
      2. If determinism is true, no one is the appropriate source of her actions.
      3. Therefore, free will is incompatible with determinism
Traditionally, compatibilists have resisted BLI and BSI by resisting the second premises of each argument. And incompatibilists have advanced BLI and BSI by offering arguments in support of the second premise of each.
   Compatibilists have had a devil of a time with their attempts to discredit the second premise of BLI. One well known attempt, suggested by Hobbes (1651/1997) and Hume (1751/1978), and later refined by Ayer (1954), Hobart (1934), Moore (1912), Schlick (1939), Smart (1963), and in more recent times Davidson (1973), was to give an analysis of the ability to do otherwise in terms of simple counterfactual conditionals. Since such counterfactuals are compatible with determinism, it would follow trivially that the ability to do otherwise is compatible with determinism. The analysis was roughly of the form 'an agent is able to do otherwise if an only if, were she to choose otherwise, she would do otherwise'. But all analyses of this form were shown by the likes of Chisholm (1964) and Lehrer (1968) to fall prey to devastating counterexamples. Consider an agent who has a phobia rendering her unable to choose to touch a red candy. As a result, she is not able to do so. Yet it is nevertheless true that, if she chose to do so, she would.
   While the traditional compatibilists were left searching for some way to demonstrate how an agent is able to do otherwise in a deterministic context, a cadre of traditional incompatibilists

---

[4] As for what deserved *blaming* amounts to, on my view (2011,forthcoming), it can be understood in terms of overt manifestations of morally reactive attitudes such as resentment and indignation, directed at the party in question.

(Ginet, 1966; van Inwagen, 1975; and Wiggins, 1973) had each developed in slightly different ways an argument that could be used in the service of advancing the second premise of BLI. The consequence argument holds that one is able to alter the necessary consequences of the past and the laws of nature only if one is able to alter that past or the laws of nature. So, if at a deterministic world an agent's act is the necessary consequence of the remote past in conjunction with the laws of nature, then that agent is able to act otherwise only if she can alter either the remote past or the laws of nature. Since no one can do either of those things (at least no one who is not a god), then if determinism is true, no one is able to do otherwise—and this just is the second premise of BLI.

Many contemporary compatibilists, who might be called *leeway compatibilists*, remain committed to refuting the second premise of BLI. This involves, first, showing why it is that the consequence argument and others like it are unsound[5], and, second, articulating an account of the ability to do otherwise that is demonstrably compatible with determinism.[6] But several other compatibilists, *source compatibilists*, have made a considerable departure from the traditional way of approaching the free will problem. Rather than resist BLI by fixing upon the second premise, they have been willing to grant that premise, at least for the sake of argument. Instead, they have focused upon rejecting the first premise of BLI, that is, upon the claim that free will requires the ability to do otherwise. [7]

Most source compatibilists who have taken up this non-traditional dialectical maneuver have done so by relying upon Harry Frankfurt's (1969) controversial argument that moral responsibility does not require the ability to do otherwise. According to Frankfurt, it is possible to construct an example in which an agent acts freely and is morally responsible for what she does but is unable to do otherwise. Frankfurt's example featured Jones, who shot Smith for his own reasons. As it happened, Black wanted Jones to shoot Smith on his own, but covertly arranged things so that if Jones were about to fail to do so, Black would manipulate Jones so that he (Jones) would shoot Smith. Since Jones acted on his own when she shot Smith, Black remained dormant; he played no role in Jones's act. Jones, it seems, acted freely and was morally responsible for doing so. Yet because of Black's presence, Jones could not have done otherwise. Hence, as the example involving Jones illustrates, moral responsibility does not require the ability to do otherwise.[8]

3. One Perspective on the Current State of the Debate

Although Frankfurt's argument remains highly controversial, many contemporary compatibilists have been convinced by some variation on it.[9] Indeed, several incompatibilists

---

[5] See, for example, Dennett (1984), Lehrer (1980), Lewis (1981), and Slote (1982).

[6] See, for example, Campbell (1997), Fara (2008), Smith (2003), and Vihvelin (2004).

[7] For example, Fischer (1994), Fischer and Ravizza (1998), Frankfurt (1971), Haji (1998), McKenna (2003), and Wallace (1994).

[8] For an examination of the further details offered in support of Frankfurt's argument, see the introduction to Widerker and McKenna (2003), and also my (2008b).

[9] Including Fischer (1994), Fischer and Ravizza (1998), Frankfurt (1971), Haji (1998), McKenna (2003), and when he's entertaining compatibilism, Mele (2006), as well as Mele and Robb (1998).

have as well.[10] These incompatibilists, who might be called *source incompatibilists*, have thus ceded BLI and instead devoted their attention to advancing BSI. What unites source compatibilists and source incompatibilists is a rejection of the leeway model of freedom, combined with an exclusive emphasis on the source model of freedom. Call those philosophers who approach the free will and moral responsibility debate in these terms *source theorists*. What divides source theorists as regards the compatibilism/incompatibilism controversy is disagreement over the second premise of BSI, that is, over the claim that if determinism is true, no one is the appropriate source of her actions.

It is in the context of the source theorists' way of understanding the dialectic that, in my estimation, the free will and moral responsibility debate largely reduces to the controversy over manipulation arguments. How so? Obviously, the second premise of BSI cannot be accepted at face value. Compatibilists will contend that they have on offer accounts of the sources of agency that are appropriate insofar as they are sufficient for the requirements of source freedom. Incompatibilists will deny this. What the incompatibilists need is some further argument in support of the second premise of BSI. To date, there are three contenders available, and to my mind, only one survives careful scrutiny. These three arguments are the ultimacy argument, the direct argument, and the manipulation argument.

The *ultimacy argument* has been defended in slightly different ways by both Galen Strawson (1986) and Saul Smilansky (2000). At its core, the ultimacy argument relies upon some version of the proposition that an agent acts freely, or instead is morally responsible for what she does, only if she contributes some necessary causal condition, N, to her action which is such that there does not exist, nor did there ever exist, any sufficient conditions for N that obtained independently of the agent herself. Call this condition the ultimacy condition. As I have argued elsewhere (2008d), in my estimation, the ultimacy condition cannot be used in an argument for an incompatibilist conclusion—which is how it is used in all variations on the ultimacy argument. It cannot be so used, since it is itself merely an expression of an incompatibilist intuition, not a consideration that ought to be used in support of incompatibilism. Hence, it might very well be true. But if true, it's a truth that cannot figure into the dispute between parties differentially committed to the compatibilism/incompatibilism debate. The incompatibilist has to use other means so as to treat this proposition as a conclusion to an argument. Interestingly, Pereboom (2001: 4) himself accepts something very much like the condition of ultimacy, but as he makes clear, he uses an argument—a manipulation argument— to argue *for* this condition. He treats this conclusion itself, the conclusion of ultimacy, as an expression of his commitment to source incompatibilism.

The *direct argument*, ingeniously developed by Peter van Inwagen (1983), is a close cousin of the consequence argument. Unlike the ultimacy argument, I believe that the direct argument poses a serious source incompatibilist challenge to compatibilism. Nevertheless, at a deep level, its credibility is dependent on something like a manipulation argument, or so I have argued (2008c). To explain, the direct argument builds on the initially compelling inference principle that one is morally responsible for the necessary consequence of a fact only if one is morally responsible for that fact. If our actions are the necessary consequences of the facts of the remote past in conjunction with the laws of nature, then according to the argument, we are responsible for our actions only if we are responsible for either the facts of the remote past or the

---

[10] See Hunt (2000), Pereboom (2001), Stump (1996), Widerker (2006), and Zagzebski (2000).

4

laws of nature. But since we're not responsible for either of those things, then under the supposition that determinism is true, we are not responsible for our actions either.

Although there are various ways that compatibilists have replied to the direct argument, I have argued that the most compelling way to do so is to resist the contention that the inference principle upon which it operates is sufficiently established that compatibilists should regard it as a threat (McKenna, 2008c: 376-7). Here is the sort of example van Inwagen has offered (1983: 187) to show that this principle, which we might call the principle of the transfer of non-responsibility (TNR), is a legitimate one:

- No one is responsible for the fact that John is bitten by a Cobra on his thirtieth birthday.
- No one is responsible for the fact that if John is bitten by a Cobra on his thirtieth birthday, then he dies.
- Therefore, no one is responsible for the fact that John dies on his thirtieth birthday.

Call this example Snakebite. The problem with using examples like Snakebite as confirming instances TNR is that they are not cases in which a person's agency is involved. The antecedent facts over which no one is responsible are not ones whose consequences include later exercises of (otherwise normal) agency; John's dying of a snakebite is nothing John does in the sense of exercising any agential capacities. Hence, citing such cases to help establish TNR's role in an argument for incompatibilism is a *non sequitur*. The trouble is, once pertinent cases are invoked, they will likely involve claims that prior sufficient conditions for which no one is morally responsible cannot give rise to actions for which an agent is morally responsible. That is, the cases highlighted will likely be cases involving typical stories about the causal histories giving rise to the actions of normally functioning agents. As might be expected, compatibilists will be prepared to claim that in candidate cases of this sort, treating them as confirming instances of TNR will just be question begging against the compatibilist.

There are, however, *some* cases in which, at least at face-value, there is theoretically-neutral intuitive weight to claiming that an agent's non-responsibility for being in a certain state, combined with the inevitability of her later action as a product of that prior state, renders her non-responsible for her later acts, despite her seeming to function as a normal human agent. These cases, however, are the kinds of cases featured centrally in manipulation arguments. Hence, I have argued, for the direct argument to get going, the incompatibilist needs confirming instances of TNR. Case like Snakebite will not do. The kinds of cases that will, however, are the ones that are at issue in manipulation arguments (McKenna, 2008c: 379-81). In these cases, what the manipulation arguments are designed to show is that the intuitions of non-responsibility that the cases are alleged to elicit can be generalized in deterministic contexts—just as a principle like TNR requires. Thus, I contend, what the direct argument needs to get up and running is a commitment to the soundness of a manipulation argument.

So, assuming the state of the dialectic is as source theorists take it to be, the debate between compatibilists and incompatibilist reduces largely to a debate over the second premise of BSI: If determinism is true, no one is the appropriate source of her actions.[11] And, as I see it,

---

[11] I say reduces 'largely' rather than 'entirely', since there are further dialectical burdens a source compatibilist shoulders even if she can establish that manipulation arguments fail. She also must provide some positive account of free action that she can show is compatible with determinism—one that nowhere implicitly relies upon an ability to do otherwise, or any assumption of indeterminism.

there is only one source incompatibilist style of argument that is currently a live option, and that is some variation on a manipulation argument. Hence, a good deal is at stake in the debate over manipulation arguments, at least from the perspective of source theorists. I turn now to an examination of manipulation arguments in general.

4. Manipulation Arguments, 'The' Manipulation Argument & Compatibilists' Options

Manipulation arguments for incompatibilism all build upon some example or other in which an agent is covertly manipulated into acquiring a psychic structure on the basis of which she performs an action. The featured agent, it is alleged, is manipulated into satisfying what might be called a Compatibilist-friendly Agential Structure (CAS), which provides conditions a compatibilist would take to be minimally sufficient for acting freely. Such an example used in the context of an argument for incompatibilism is meant to elicit the intuition that, due to the pervasiveness of the manipulation, the agent does not act freely and is not morally responsible for what she does. It is then claimed that any agent's coming to be in the same psychic state through a deterministic process is no different in any relevant respect from the pertinent manner of manipulation. Hence, it is concluded that CAS is inadequate. Often, it is then suggested, or more explicitly argued for directly, that no candidate CAS will turn the trick; free will and moral responsibility are incompatible with determinism, full stop.

To illustrate, consider Richard Taylor's formulation of a manipulation argument. Taylor targeted a classical compatibilist account of freedom according to which one is free so long as her volitions are not impeded in acting as she does. About such a view, he wrote:

> But if the fulfillment of these conditions renders my behavior free—that is to say, if my behavior satisfies the conditions of free action set forth in the theory of soft determinism—then my behavior will be no less free if we assume further conditions that are perfectly consistent with those already satisfied.
>
> We may suppose further, accordingly, that while my behavior is entirely in accordance with my own volitions, and thus "free" in terms of the conception of freedom we are examining, my volitions themselves are caused. To make this graphic, we can suppose that an ingenious physiologist can induce in me any volition he pleases by pushing various buttons on an instrument to which, let us suppose, I am attached by numerous wires. All of the volitions I have in that situation are, accordingly, precisely the ones he gives me. By pushing one button, he evokes in me the volition to raise my hand; and my hand, being unimpeded, rises in response to that volition. By pushing another, he induces the volition in me to kick, and my foot, being unimpeded, kicks in response to that volition. We can even suppose that the physiologist puts a rifle in my hands, aims it at a passerby, and then, by pushing the proper button, evokes in me the volition to squeeze my finger against the trigger, whereupon the passerby falls dead of a bullet wound.
>
> This is the description of a man who is acting in accord with his inner volitions, a man whose body is unimpeded and unconstrained in its motions, these motions being the effects of his inner states. It is hardly the description of a free and responsible agent. It is the perfect description of a puppet. (1963: 45)

Note that the particular compatibilist account of CAS that Taylor was aiming to discredit falls far short of the more robust theories currently on offer, but his strategy is suggestive of how more recent proposals could be attacked.

To offer yet another illustration of a manipulation argument, consider instead Robert Kane's formulation. Kane appeals to the story of Skinner's *Walden Two* as a way of discrediting Frankfurt's proposed account of CAS:

>We are well aware of these two ways to get others to do our bidding in everyday life. We may force them to do what we want by coercing or constraining them against their will, which is constraining or CC control. Or we may manipulate them into doing what we want while making them feel that they have made up their own minds and are acting "of their own free wills"—which is covert nonconstraining or CNC control. Cases of CNC control in larger settings are provided by examples of behavioral engineering such as we find in utopian works like Aldous Huxley's *Brave New World* or B.F. Skinner's *Walden Two*. Frazier, the fictional founder of Skinner's Walden Two, gives a clear description of CNC control when he says that in his community persons can do whatever they want or choose, but they have been conditioned since childhood to want and choose only what they can have or do.
>
>Utopian versions of CNC control such as Skinner's are an excellent medium for expressing [an] objection to hierarchical views of free will such as Frankfurt's. For the citizens of Walden Two, as they are described, are marvelously "wholehearted" in their attitudes and engagements. They are "satisfied" with themselves (almost to a cloying degree) and "have the wills they want to have." As a result, Frazier, the founder of Walden Two, insists that it is "the freest place on earth." And why should it not be? There is no coercion in Walden Two and no punishment. All citizens can have and do whatever they want or choose; and they can *will* to do whatever they *want*—that is, their first-order desires always conform to their second-order volitions. They have maximal freedom of will and action in the hierarchical sense.
>
>Yet most people who look at Walden Two would say—and say rightly, in my opinion—that its citizens lack free will in a deeper sense than being able to do what they want and will what they want. In this deeper sense, their wills are not "their own" because they are not the original creators of their own ends or purposes. Their ends or purposes are created by their conditioners or controllers. (1996: 65)

Kane does not take the step here of generalizing in a way that would apply to all extant forms of compatibilism, and so to a general conclusion that incompatibilism is true. He restricts his remarks to hierarchical models. But it is easy to see how his argument could be extended.

Thus far I have written in terms of manipulation arguments—plural. This is to make perspicuous that there are different instances of a general form. Let's refer to the argument form as *the* manipulation argument. So, treating 'X' as a placeholder for different ways of manipulating an agent, consider *the manipulation argument (TMA)*:

1. If an agent, S, is manipulated in manner X to perform act A from CAS, then S does not A freely and is therefore not morally responsible for A-ing.
2. Any agent manipulated in manner X to A is no different in any relevant respect from any normally functioning agent determined to do A from CAS.
3. Therefore, any normally functioning agent determined to do A from CAS does not A freely and therefore is not morally responsible for A-ing.

Different instances of TMA will arise in relation to: (i) different candidate cases of manipulation, and (ii) different compatibilist proposals for CAS.[12]

---

[12] TMA is the only argument form I will consider. But note that Mele (2008) distinguishes three different forms. The first, which he calls a *straight manipulation argument*, is basically the form I have set out. The second form, a

How ought compatibilists respond to TMA, and to the various particular versions of manipulation arguments? Given the variability of (i) and (ii), there is no one-size-fits-all compatibilist reply to the full range of manipulation arguments. Whether a compatibilist is positioned to resist either the first or second premise of a particular manipulation argument will depend on the details of the manipulation case and on the content of her account of CAS. To reject the first premise of a manipulation argument is to adopt a *hard-line reply*, since it involves taking on directly the hard burden of resisting the (often) intuitively plausible judgment that a suitably manipulated agent does not act freely and is not morally responsible. To reject the second premise of a manipulation is, by contrast, to adopt a *soft-line reply*, since it permits that an agent manipulated in the way featured by the argument does not act freely and is not morally responsible. Instead, it involves showing how it is that an agent who is merely determined to act as she does differs in some relevant way from an agent so manipulated.

Despite that fact that there is no one-size-fits all reply to manipulation arguments, in my estimation, compatibilists have good reason to adopt a general policy with respect to them. When reflecting upon the manipulation argument—that is, TMA, the argument form—I have argued that compatibilists should adopt a policy of inclining toward a hard-line reply to various instances when doing so is a viable option (McKenna, 2004: 216-7; 2008: 143-4). This, however, seems counter-intuitive. Hard-line replies apparently involve bullet-biting. Soft-line replies allow intuitions about freedom-destroying manipulation cases to go undisturbed. So why advise compatibilists to incline toward hard-line replies? Why look for bullets to bite? The reason is simple. Compatibilists can win lots of battles by adopting a soft-line reply, pointing out how their respective accounts of CAS are not satisfied by various cases of manipulation. They can safely win these battles by allowing that the manipulated agents in question are not free and responsible, pointing out how it is that the agents so manipulated are different from what is involved in being a free agent at a determined world. But the trouble with this dialectical maneuver is that it is often open to the incompatibilist to parry with a simple adjustment to her case of manipulation, one that accommodates the compatibilist in such a way that the compatibilist's preferred account of CAS *is* satisfied. Thus, in adopting a soft-line reply to pertinent instances, the compatibilist just forestalls the inevitable, which is a revised case of manipulation, and a new instance of a manipulation argument for which she'll not have the option of a soft-line reply—for which, she'll *have* to face the music and take on directly a hard-line reply.

There is only one way for the compatibilist to resist the claim that she'll inevitably be forced to take a hard-line reply in response to some version of a carefully crafted manipulation argument. She must contend that there is a principled reason why it is that the deterministic causes issuing in a free act cannot be artificially replicated. I see no convincing way to rule this out as a metaphysical or at the very least a conceptual possibility. Hence, I have advised

manipulation argument to the best explanation, replaces the second premise as I have set it out above with a best-explanation premise of roughly the form, 'the best explanation that S is not free and morally responsible when manipulated in manner X into A-ing is that S is causally determined'. The third form, an *original design argument* of the sort Mele himself has developed (2006), works not from a case of an already existing agent who is manipulated, but from an agent who is originally designed to perform some act, A. Mele is correct to mark these distinctions. However, it is not clear to me how much the distinction between the first and third comes to. Originally designing an agent to do something at a later time seems also to be a case of manipulation, but a case in which the agent is manipulated from a point in time prior to her coming into existence. Kearns (forthcoming) makes a similar point.

compatibilists to cut to the chase (2008: 144-5). If an incompatibilist offers a manipulation argument and her featured case of manipulation is just off the mark of satisfying a compatibilist's preferred account of CAS, I think the compatibilist is better off taking on the case under consideration. Rather than take a soft-line reply, she should offer whatever slight amendment to the case is needed in order to get her account of CAS just right. Once she's fixed up the case, she should then go straight for a hard-line reply.

Of course, it will not *always* be the case that a compatibilist should make such a move. I've only advised compatibilists to adopt a general policy of favoring a hard-line reply when doing so is a viable option. Sometimes the case of manipulation will be far enough off target that the compatibilist will be easily suited to point out that her account of CAS does not presume that all free agents are anything like that. But for that range of cases where the incompatibilist is hot on the trail of getting at a compatibilist's account of CAS with a well crafted manipulation case, the compatibilist should just commit to the hard-line.

5. Pereboom's Four-Case Manipulation Argument

To build his manipulation argument, Pereboom attacks a version of CAS that is a conjunction of four influential contemporary compatibilist proposals: Ayer's, Frankfurt's, Fischer and Ravizza's, and Wallace's. For ease of discussion, let us refer to this conjunctive account as 'CAS*'. Given CAS*, the manipulated agent featured in Pereboom's argument is not constrained by any irresistible desires and does not act out of character (Ayer, 1954); he acts upon first-order desires with which he identifies at a high-order (Frankfurt, 1971); he is responsive to reasons in a way that displays a stable, sane pattern (Fischer and Ravizza, 1998); and when he acts, he has the general capacity to regulate his behavior in light of moral reasons (Wallace, 1994). Although he does not say so explicitly, Pereboom clearly intends to target more than extant compatibilist proposals; all compatibilist comers are open to the same treatment. Whatever further compatibilist proposals one might come up with could just be conjoined to CAS*.

What makes Pereboom's argument especially compelling, and, to my mind, marks a genuine innovation in the debate, is that he does not just work from one manipulation case. Rather, he employs a series of them, three, in such a way that the first in the series involves a kind of manipulation that is very different from the kind of casual process that would be expected to unfold at a deterministic world of the sort that ours might (could?) be. He then marches through two more cases so that each case comes closer and closer to a case in which an agent acts as she does in a world at which determinism is true. In doing so, he employs a generalization strategy of treating like cases alike, and in transition between the cases, argues that there is no relevant difference between the modes of manipulation, so that if we are to treat the first case as one in which the manipulated agent is not free and responsible, that will carry over to the next in the series, and so forth, with the fourth and last merely being a case in which determinism is true. This generalization strategy is used to back up the second premise in Pereboom's argument—that is, it is meant to show that the highlighted means of manipulation, even in an extreme case such as his first, are no different in any relevant respect than is the means of causation at work under the assumption of determinism.

Here is a truncated summary of each case:

> *Case 1*: Professor Plum was created by a team of neuroscientist, who can
> manipulate him directly through radio-like technology, but he is as much like an
> ordinary human being as is possible, given his history. The scientists "locally"

> manipulate him to undertake a process of reasoning, directly producing his every
> state moment by moment, which leads to the killing of Ms. White for egoistic
> reasons. (Pereboom, 2001: 112-3)

The manipulation is such that Plum fully satisfies CAS* when he murders Ms. White. Of course, we are to have the intuition that, due to the extremity and pervasiveness of the manipulation, Plum does not act freely and is not morally responsible for killing Ms. White.

> *Case 2*: Plum is like an ordinary human being, expect he was created by a team of
> neuroscientists who, although they cannot control him directly, have programmed
> him to weigh reasons for action so that he is often but not exclusively egoistic,
> with the result that in the circumstances he is casually determined to undertake the
> process that results in his killing Ms. White. (Pereboom, 2001: 113-4)

Again, Plum satisfies CAS*. As for the time lag between the neuroscientists' manipulations and Plum's act, there is no relevant difference, Pereboom argues, between this case and Case 1 where the manipulation is direct and moment by moment. Treating like cases alike, we are to have the intuition that Plum does not act freely and is not morally responsible for killing Ms. White.

> *Case 3*: Plum 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… process… that results in his killing Ms.
> White. (Pereboom, 2001: 114)

Yet again, Plum satisfies CAS*. Case 3 is a common sort of case that we would find in the actual world. But the causal inputs, Pereboom reasons, are just less weird causes as in comparison with the sort highlighted in Case 2. Otherwise, there is no relevant difference. So, treating like cases alike, we are to have the intuition that Plum does not act freely and is not morally responsible for killing Ms. White.

> *Case 4*: Physicalist determinism is true, and Plum 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 Ms. White comes about as a result of his undertaking the… [relevant] process.
> ([my braces], 2001, p.115)

Once again, Plum satisfies CAS*. And since Case 4 differs from Case 3 merely by virtue of the fact the deterministic causes apply universally rather than locally as regards the details of Plum's upbringing, there again seems to be no relevant difference between the cases. So, Pereboom argues, treating like cases alike, we should conclude that Plum in Case 4 does not act freely and is not morally responsible for killing Ms. White. Ergo, CAS* is refuted.

According to Pereboom, by way of his manipulation argument and his march from Cases 1 through 4, he is able to exploit Spinoza's strategy of making salient the hidden causes of human actions so as to show how, once they are revealed, the illusion of control dissipates:

> Men think themselves free, because they are conscious of their volitions and appetite, and
> do not think, even in their dreams, of the causes by which they are disposed to wanting
> and willing, because they are ignorant of [those causes]. (Spinoza, 1677/1985, v. 1, 440,
> as quoted in Pereboom, 2008: 161)

Seeing that being causally determined (Case 4) really is just like, in all relevant respects, being manipulated moment by moment by a team of neuroscientists (Case 1) helps crystallize the incompatibilist intuition that determinism destroys free and responsible agency.

6. A Hard-line Reply

Several critics have opted for a soft-line reply to Pereboom. Worries about his first two cases have led some to think that the manipulated agents in them fall far short of satisfying anything like CAS*. For instance, some have worried that in Case 1, with Plum manipulated moment by moment, Plum's agency is bypassed altogether (Demetriou, 2010: 601-5; Fischer, 2004: 156; Haji, 1998: 24; and Mele, 2005: 78). And others have worried that in Case 2, Plum's programming is such that he does not have the kind of history a suitably sensitive historical compatibilist would find acceptable (Haji, 2009: 166-8; and Mele, 2005: 78). Admittedly, Pereboom did under-describe these cases, and it is understandable that some of his critics have interpreted him in these ways. Nevertheless, I have argued that there are charitable ways to understand Pereboom's cases so that, for instance, Plum's agency in Case 1 is not subverted but rather constituted by the modes of manipulation, and in Case 2 Plum was manipulated in such a way that he *did* undergo a compatibilist-friendly history (McKenna, 2008: 148-52).[13] In doing so, I've helped Pereboom along with his cases. I'll not rehearse the details here. For present purposes, grant that they are satisfied in ways that entitle Pereboom to claim that CAS* is satisfied in his sequence of cases 1 through 3 (as are any further compatibilist-friendly historical conditions one might think fit to add). Hence, grant that a soft-line reply is off the table.

Turning to a hard-line reply, I shall assume that a proper reply to Pereboom requires resisting the contention that Plum in Case 1 does not act freely and is not morally responsible. To achieve this result, in an earlier piece (2008), I first focused attention on pertinent agential and moral properties of Plum *as displayed in Case 4*. In this case, Plum, we can easily imagine, is just as much like an ordinary moral agent as any one of us. His phenomenology is just as sophisticated. He is capable of feeling incredible remorse, or instead ambivalence about killing Ms. White, or alternatively, delicious pride. Furthermore, we can suppose that Plum had a rich history of moral development just like any psychologically healthy person who emerges from childhood into adulthood. And we can also allow that Plum is an agent who lives up to the kind of emotional complexity of the sort Strawsonians highlight. He is capable of resentment and gratitude, moral indignation and approbation, guilt and pride, and he is richly sensitive to the significance of these morally reactive emotions in his relations with others. He is, in this way, a full-fledged member of the moral community, one whose standing enables the kinds of interpersonal relations out of which adult moral life arises. The point of attending to these details is to help elicit the intuitions that are friendly to compatibilists—to bring forth the sense that a determined agent of the sort Plum is in Case 4 is a richly complex agent just like any person we might come across in the course of our daily lives.

Second, I sought to clarify the initial attitude both compatibilists and incompatibilists are entitled to adopt toward Plum in Case 4. Naturally, the compatibilist is not entitled to insist that it be granted that Plum in Case 4 *does* act freely and *is* morally responsible. That is precisely what is in dispute in considering the manipulation argument currently on the table. Nor, for this

---

[13] And Pereboom (2005) has responded to these objections just as I have predicted that a careful incompatibilist would. He's clarified or modified his cases so that, according to him, they do satisfy the agency conditions that those opting for the soft-line have fixed upon.

reason, should the compatibilist expect that the intuitions that she seeks to elicit in Case 4—by focusing on salient agential and moral properties—speak commandingly in favor of compatibilism. All the compatibilist can expect is that an open-minded inquirer, one who is undecided and so agnostic by the prospect that determinism rules out freedom and responsibility, could be moved to see the plausibility of compatibilism. But equally to the point, incompatibilists cannot assume at the outset that Plum in Case 4 is not free and is not morally responsible, nor ought they to expect that the initial intuition about Plum in Case 4 is that, clearly, he does not act freely and is not morally responsible. Again, that is what is in dispute, and assuming this at the outset would be patently question-begging. Thus, the stance towards Case 4 that an incompatibilist such as Pereboom is dialectically obligated to take in executing his argument is the same one that the compatibilist must take. Both must allow that in Case 4, it is an open question whether Plum acts freely and is morally responsible for killing Ms. White.

Third, armed with the two preceding considerations, I have attempted to turn Pereboom's strategy against him. Pereboom begins with his Case 1 and, seeking to elicit an incompatibilist intuition in response to the disturbing manipulation featured in it, marches through to his Case 4, thereby drawing an incompatibilist conclusion. I grant that Pereboom does gain *some* intuitive advantage in favor of incompatibilism by calling attention in Spinozistic fashion to hidden causes revealed in Case 1 and transferring through to Case 4. That, I acknowledge, has to be weighed on the scales when evaluating the overall strength of his argument. Nevertheless, I propose that the compatibilist move in the opposite direction, and that this too must be weighed in the scales. Begin with Pereboom's Case 4. By calling attention to the salient agential and moral properties of Plum in Case 4, I have sought to elicit the intuition amongst an audience of undecided, open-minded inquirers that it is not clearly the case that determinism rules out free will and moral responsibility. It is, at least, a plausible contention. Since, by Pereboom's own lights, Case 3 differs in no relevant respect from Case 4, and since we are also instructed to treat like cases alike, the compatibilist should argue that we should treat Case 3 just as we treat Case 4. Likewise we should treat Case 2 no differently than we treat Case 3, just as Pereboom contends, and also, finally, Case 1 no differently than we treat Case 2. Hence, we get the result that we should treat Case 1 no differently than we treat Case 4. [14]

To clarify: how, exactly, should we treat Case 1, according to this hard-line compatibilist strategy? We should transfer the initial undecided open-minded attitude of agnosticism that perhaps Plum in Case 4 does act freely and is morally responsible. That is, given this strategy, and the initial dialectical attitude the compatibilist is entitled to in Case 4, the compatibilist is in no position to conclude that in Case 1 Plum *does* act freely and *is* morally responsible. She's only entitled to conclude that it is not clear that Plum does not act freely and is not morally responsible. And she is entitled to conclude this much by way of carrying over the intuitive weight regarding Case 4 that she seeks initially to elicit from an undecided audience by inviting them to focus carefully on the rich agential and moral properties that Plum in Case 4 has—all of which, it can be granted, carry over just as much to Plum in Cases 3, 2, and 1. [15]

---

[14] Fischer (2011) has recently responded to another manipulation argument, Mele's (2006) zygote argument, in a similar fashion.

[15] To heighten the effect of my strategy, I actually added two cases to Pereboom's four cases so as to make the transition from a mere determined agent to Plum in Pereboom's Case 1 even smoother. I'll set that detail aside here, since nothing in my current response to Pereboom will turn on it.

Of course, it will not go unnoticed that my proposed compatibilist hard-line response does not require taking the bold position of maintaining that in Case 1 Plum does act freely and is morally responsible. Some will likely regard this as a copout. I do not think it is. The dialectical burden of the compatibilist who adopts the hard-line in response to Pereboom is simply to discredit the first premise in this manipulation argument. And this is achieved, I contend, merely by casting sufficient doubt on the truth of the premise. It's the incompatibilist who has offered the argument, and it's he who is claiming that intuitions about a case like Case 1 will speak commandingly in favor of the proposition that the featured agent does not act freely and is not morally responsible. It's not the compatibilists' burden in this context to convince fair and open-minded inquirers that clearly an agent manipulated the way Plum is in Case 1 does act freely and is morally responsible. This crowd—the audience of fair and open-minded inquirers—was not even expected to have that intuition about Case 4 antecedent to reflecting on the more outrageous cases like Cases 1 and 2.

To help cement this hard-line compatibilist response to Pereboom, I also took one further step to help lessen the jarring counter-intuitive suggestion that, perhaps, an agent like Plum in Case 1 does act freely and is morally responsible. The central idea is that there are actual cases which are in many respects just like the imaginary outlandish manipulation cases featured in an argument like Pereboom's, and these cases are not all that uncommon. Here, our intuitions do not seem to favor an incompatibilist diagnosis. As Nomy Arpaly has thoughtfully pointed out when reflecting on the nature of different manipulation cases, there are all sorts of real life cases where persons undergo spectacular changes as a result of things like religious conversions, the birth of a child, one's hormones drying up, and so on (Arpaly, 2006: 112-3). Indeed, there's no end to the number of naturally occurring real-life cases of "manipulation" in which massive and unexpected alterations disrupt people's lives in ways that dramatically reconfigure their psychic constitutions: people suffer traumatic accidents; have a loved one die in their arms unexpectedly; are crushed during their youths by the weight of violent parents. And yet, if these types of changes leave the adult person otherwise, sane, rational and stable, in ways that would allow satisfaction of pertinent compatibilist conditions on free action, very few are inclined to think that these kinds of "manipulation" cases undermine a person's free agency and her responsibility.[16] They are just regarded as the (often) tragic contingencies out of which these people manage to fashion their moral personalities and thereby come to be the kinds of agents they are.

7. Pereboom's Resistance to the Hard-line Reply

To resist my hard-line reply, Pereboom challenges my contention that an initial agnosticism towards Case 4 should transfer through to Case 1. Rather, he argues, if we get clear on the dialectically appropriate state that should be expected of those whose intuitions ought to

---

[16] Mele (1006: 179-84) has objected to the conclusions Arpaly draws from these cases. According to him, when such changes are freedom-and-responsibility preserving, they permit a process whereby an agent is able to accept or instead reject them freely. Of course, if this were so, these cases would not be relevantly like manipulation cases after all. Here, I would just voice my disagreement with Mele. Some cases might very well be as Mele describes them, but I agree with Arpaly that it is a pervasive fact about the human condition that persons sometimes undergo such changes through processes that bypass altogether any capacities to freely accept or reject the changes. And in many of these cases, these persons are *still* regarded as free and responsible agents with respect to those actions issuing from the pertinent changes.

be informative as regards his argument, we shall see that cases like his Case 1 and Case 2 should still have the effect of recommending an incompatibilist diagnosis as regards Case 4.

One point of agreement between Pereboom and me has to do with those whose intuitions are *not* the target of a manipulation argument (Pereboom, 2008: 160-3). There are some who, upon hearing about the puzzle regarding freedom and determinism are, so to speak, natural incompatibilists; their untutored intuitions are such that they need not be prompted by reflection upon manipulation cases and a corresponding argument, or for that matter any other sort of incompatibilist argument, to adopt an incompatibilist posture. There are others who, upon learning of the puzzle about free will are instead natural compatibilists; their untutored reactions are to think that determinism does not pose much of a threat to freedom and responsibility at all, and they do not need to be prompted by further arguments on behalf of compatibilism. In either of these groups—those of the natural incompatibilists as well as the natural compatibilists— intuitions about responses to manipulation cases will not prove fruitful for settling the dispute between compatibilists and incompatibilists. Natural incompatibilists can be expected to respond to pertinent manipulation cases in ways that predictably confirm their predisposition. And natural compatibilists can be expected to do likewise. (Or, instead, if natural compatibilists do respond differently to learning of relevant manipulation cases, it is likely that they will be inclined to retain the conviction that there must be some relevant difference between determinism and such manipulation, even if there is difficulty stating what that comes to.) So the turf over which incompatibilists and compatibilists are doing battle with respect to manipulation arguments like Pereboom's four-case argument is the turf of the undecided.[17] But how are we to understand the appropriate *initial* epistemic stance of this undecided group in relation to the stance they ought to adopt *after* reflecting upon cases like Pereboom's Case 1 and Case 2?

According to Pereboom, my implicit answer to this question shows where I have gone wrong, and why it is that he is entitled to claim that cases like his Case 1 and 2 can be counted on to generate in the appropriate audience an incompatibilist conclusion. Here is how Pereboom formulates my position:

> [McKenna's] central idea is that whatever attitude it is rational initially to have about Plum's responsibility in the ordinary deterministic examples (my [Pereboom's] case 4…) transfers to the manipulation cases, since, after all, the incompatibilist has set up the cases so that there are no responsibility-relevant differences among them. Because it is at the outset rational for us to have an agnostic attitude about the claim that Plum is morally responsible in the ordinary deterministic examples, the absence of relevant differences allows the rational agnosticism to transfer unimpeded to the manipulation cases, thereby depriving them of counting in favor of incompatibilism. (2008: [my braces] 160)

---

[17] Note that Mele frames the dialectic in a different manner than Pereboom and I do. Mele certainly agrees with us that some respondents to manipulation cases cannot be counted on to have especially informative intuitions insofar as their reactions to pertinent cases are in one manner or another already theoretically loaded (2006: 192). But where he differs from us is with respect to those who might be called "untutored subjects" (190). Here, Mele thinks these subjects' intuitions are not helpful and that it is more useful to consult the reflective judgments of those who "have thought long and hard about freedom and responsibility and are agnostics about compatibilism" (190). Mele, who is one such philosopher (actually, the only one I know of), reports that he is not inclined to come down in favor of compatibilism (194) in a way that would involve the assertion that the featured agent does indeed act freely. But Mele *also* does not claim that he is convinced that pertinently manipulated (or created) agents are not free and responsible. As I see it, this counts as a win for the compatibilist, insofar as it indicates by a means different than the one I employ here that an especially provocative manipulation (creation) argument—Mele's Zygote Argument— is *not* decisive. As a compatibilist responding to an incompatibilist's argument, I'll settle for credible doubt.

Pereboom then attributes to me the following proposition as it applies to the initial attitude one ought to adopt towards his Case 4: (M) *It is not clear that Plum is not morally responsible* (161). It is this attitude, (M), that I contend is a rational one to adopt as transferring through to cases 3 through 1. My mistake, according to Pereboom, is in failing to appreciate how it is that cases like his Case 1 and Case 2 could count as clarifying considerations that ought to lead the targeted audience to reject (M) in favor of an incompatibilist diagnosis.

To make his point perspicuous, Pereboom distinguishes between different "undecided" stances. One stance is the stance of the *confirmed agnostic*, who, upon initial reflection as regards the puzzle about freedom and determinism, is resolute in her agnosticism and believes that no further enquiry could settle the matter (162). As Pereboom points out, this audience could be expected to transfer (M) from a case like Case 4 through to Case 1; this is what they rationally ought to do given their initial attitude. If this were the dialectically appropriate audience to target, Pereboom, notes, then my reply to him would be warranted (163). But Pereboom rightly rejects this as the appropriate initial attitude. According to him, the appropriate approach:

> …affirms that determinism provides a reason for giving up the responsibility assumption, but claims that so far the issue has not been settled. Its advocate would say about an ordinary case of an immoral action, in which it is specified that the action results from a causally deterministic process that traces back beyond the agent's control, that it is now in question whether the agent is morally responsible. Call this the *neutral inquiring* response. By this response it is initially epistemically rational not to believe that that the agent in an ordinary deterministic example is morally responsible, and not to believe that he is not morally responsible, but to be open to clarifying considerations that would make one or the other of these beliefs rational. (162)[18]

So, according to Pereboom, the appropriate initial response towards his Case 4 by members of the pertinent target audience is the *neutral inquiring response* (NIR). But what kind of "clarifying considerations" would be of use for such an audience? Pereboom writes:

> …adducing an analogy in which one's intuitions are clearer might itself count as the relevant sort of clarifying consideration. In a situation in which the neutral inquiring response to an ordinary determined agent is at first epistemically rational, it may be that an analogous manipulation case functions as a clarifying consideration that makes rational the belief that the ordinary determined agent is not morally responsible. (162)

Of course, as Pereboom sees it, intuitions about a case like Case 1 or Case 2 are clearly friendly to incompatibilism, and these cases provide the right sort of analogy to Case 4 to help clarify the rationality of one's beliefs.

So, in short, according to Pereboom, my contention that (M) transfers through from Cases 4 through 1 rests on the fallacious assumption that the proper initial attitude towards Case 4 is that of the confirmed agnostic, not that of the open-minded and undecided neutral inquirer. Once the appropriate initial attitude is clarified, it can be seen that (M) will not—and, more importantly, rationally ought not—comfortably transfer through from Case 4 to Case 1. It ought

---

[18] Note that Pereboom's formulation here is probably a bit too restricted. There are also those who would *not* affirm that determinism provides a reason for giving up the responsibility assumption, but who *also* claim that the issue has not yet been settled. Daniel Haas (unpublished) has recently developed this compatibilist-friendly insight in response to Pereboom.

rather to collapse under the weight of the intuitive force of cases like Case 1 and Case 2 as these will clarify what really is going on, even in an ordinary case like Case 4.

Building upon the preceding indictment, Pereboom entertains the possibility that a compatibilist might wish to contend that the neutral inquiring response initially adopted toward Case 4 ought to remain the appropriate attitude to adopt even upon confronting cases like his Case 1 and 2. If so, these cases would not after all count as clarifying considerations on behalf of incompatibilism (163-4). This could then be used by the compatibilist to place in check Pereboom's reply to me. But here Pereboom correctly points out that this manner of resisting him would commit a compatibilist to the implausible claim that even initially upon facing a case like Case 1 or Case 2, there is no preliminary intuition of non-freedom and non-responsibility that such cases can be expected to elicit (164). This, he points out, would amount to a departure from my own position. Indeed, it would. How so? The strategy I adopted of transferring one's response toward Case 4 through to Case 1 was meant to function as a tonic to the expected initial reaction to Case 1 that, at least provisionally, an agent so manipulated is, intuitively, thought to be such that she does not act freely and is not morally responsible.

8. A Hard-liner Takes it on the Chin

Pereboom's fair-minded response offers a constructive and insightful way to move the debate further along. But I do not think it is adequate to block my hard-line reply to his original four-case manipulation argument. I am in complete agreement with Pereboom as regards the initial dialectical stance that both compatibilists and incompatibilists should adopt when approaching his manipulation argument—or, for that matter, any manipulation argument. Both compatibilists and incompatibilists need to suppose that the neutral inquiring response towards ordinary determined agents (of the sort featured in Pereboom's Case 4) is initially epistemically rational. However, from this supposition, it would be wrong to infer that, short of admitting defeat, it is not open to a hard-liner to grant that manipulation cases like his Case 1 and Case 2 can count as clarifying considerations speaking on behalf of incompatibilism. Thinking this concession off limits, one might then assume that a hard-liner is fallaciously driven to adopt the stance of the confirmed agnostic. Otherwise, it might be thought, how can the hard-liner argue that the initial attitude (M) does after all transfer from Case 4 to Case 1? Won't the effects of Case 1 and Case 2 undermine (M)?

Here is the key inference that I wish to resist: If manipulation cases like his Case 1 and Case 2 provide intuitive weight in favor of a judgment of non-freedom and non-responsibility, then they count as clarifying considerations that are sufficiently strong that they ought rationally to move one off of a neutral inquiring stance as regards a case like his Case 4. It is this assumption that I wish to challenge.[19] I grant that cases like Pereboom's Case 1 and Case 2 *do* initially and presumptively weigh in the direction of judging that the Plums in each of them do not act freely and are not morally responsible. And I grant that, to this extent, they count as incompatibilist-friendly clarifying considerations as regards the initial neutral inquiring response toward Case 4. In making these concessions, I am, so to speak, taking it on the chin. Pereboom's argument is a powerful one, and it would be nothing shy of intellectually dishonest not to recognize it as such. But what I wish to challenge is that his Cases 1 and 2 are *sufficiently* compelling to move one fully off the neutral inquiring response. Speaking in terms of degrees of

---

[19] To avoid any misimpressions, I am *not* attributing this inference to Pereboom. It's just that, in making clear how I distance myself from it, I can bring into relief how it is that a hard-liner like me should resist Pereboom.

credence, cases like Pereboom's Case 1 and 2 might help move the neutral inquirer to raise her credence somewhat in favor of a judgment of incompatibilism as regards Case 4, and to that extent be clarifying, but, I shall now argue, there is reason to believe that such manipulation cases ought not be regarded as decisive. To do so, I offer the following three points.

First, Pereboom wishes to make clear that cases like his Case 1 and Case 2 can count as clarifying considerations that ought to move relevant parties in the direction of an incompatibilist diagnosis. But equally to the point, consider those whose initial attitude toward a case like Case 1 or Case 2 is "Clearly, certainly, obviously, such an agent is not free and is not morally responsible!" Calling attention to the way in which these outrageously manipulated agents could be just like a normal person any one of us might meet (a Case 4), but with very strange or unexpected causal strings, *also* gets to count as a clarifying consideration. And this should provide some reason for those who would adopt a neutral inquiring response toward a case like Case 4 to at least consider looking askance at their initial intuitive reaction to such cases as Case 1 and Case 2. It is for this reason that, in advancing a hard-line reply, I have argued that the compatibilist needs to fill out the manipulation case as clearly as possible, so as to make clear that the candidate for manipulation is not made out to appear no more than, say, a puppet or a zombie—a mere comic book sketch of a full-blooded person.[20] The modes of manipulation have to be such that they allow for an agent whose complexity and psychic depth are no different from any one of us. And, I contend, in advancing a hard-line reply, the compatibilist needs to make this vividly clear to her target audience of neutral and open-minded inquirers.

Second, while cases like Pereboom's Case 1 and Case 2 are to be placed on the scales in weighing the intuitive force of incompatibilism, so too should the kinds of cases to which Arpaly has called attention. These cases, I have noted, are unfriendly to an incompatibilist diagnosis. And they too get to count as clarifying considerations—considerations that, I say, counterbalance at least to some degree the force of cases like Case 1 and Case 2. What's especially noteworthy about these kinds of cases is that they are all akin to Pereboom's Case 3, wherein Plum is simply raised by parents in an environment that has powerful influence on Plum's values as he emerges into adulthood. In my earlier piece, to bring a case like this to life, I drew upon the example of Ann, a woman whose early childhood was shaped by the protracted death of a parent suffering through chemotherapy when that form of treatment was still in its infancy, when it was just plain barbaric (2008: 256-7). Ann, I explained, came to value life in a certain way from those early experiences—as something that should not be squandered, that should be lived to its fullest, with no promise of a long future or a lovely afterlife. And so her life was shaped. Ann, I noted, did not regard these facts about her history as impediments to her freedom, but instead as sources of it, as conditions of her agency and sources of her autonomy. The case of Ann, I proposed, just is a real life "manipulation" case, but for the fact that it was not intentionally brought about by another agent for the purpose of so shaping Ann. For the most part, the most common, natural, intuitive response to these kinds of cases is to persist in regarding such agents as free and responsible.

In his reply, Pereboom thoughtfully countenances my case of Ann (Pereboom, 2008: 167). He notes that he has available the conceptual resources to acknowledge the value of Ann's

---

[20] Recall Taylor's formulation of a manipulation argument as quoted above. In it, he describes the freedom-undermining nature of his manipulation case as a perfect description of a puppet. The compatibilist should not allow this to pass. Puppets have no psychic depth of any sort, and are at best caricatures of what a full-blooded human agent is like.

life and its direction despite his disinclination to think of her as rendered free and responsible by her history. Her virtuous achievement, Pereboom remarks, can be enjoyed and celebrated (167). Furthermore, Pereboom is careful to acknowledge that cases like the Ann case *do* count in favor of a compatibilist diagnosis and against his own position. Indeed, he regards it as a cost of his position that it is not in keeping with the likely widespread response that a person like Ann is free and responsible with respect to her actions issuing from the relevant values. Pereboom's remarks here strike me as especially charitable and fair-minded. But in reply, I would note that it is illuminating that he seeks alternative means to account for cases like the Ann case—a case very much in the family of cases like his Case 3. Why, oughtn't the case of Ann just be treated as a further clarifying consideration as regards the neutral inquiring response, a clarifying consideration that, in assessing (M), competes for credence with responses to cases like his Case 1 and Case 2? In fact, one might wonder whether cases like the case of Ann should be afforded a *greater* degree of significance in this dispute between compatibilists and incompatibilists. It is to this point that I shall now turn.

Third, it is worth examining the relative status of intuitions about cases like Pereboom's Case 1 and Case 2, as in comparison with intuitions about cases like the Arpaly cases and my case of Ann. It is one thing to reflect upon how those untutored subjects who are open-mined, neutral inquirers both would and ought rationally to respond to *all* of these cases: What grounds do these cases offer as a basis for moving this audience either off of (M) and in the direction of incompatibilism, or instead more confidently toward (M) and thus away from incompatibilism? But it is another for philosophers engaged in the free will dispute to assign these different responses from the pertinent audience their relative overall significance in assessing an argument like Pereboom's four-case argument: How ought we as philosophers assess what the crowd of neutral inquirers both would and ought rationally to do?

I propose that philosophers assessing manipulation arguments should assign responses from the neutral inquiring audience to cases like my case of Ann and the Arpaly cases *a greater degree of clout* as in comparison with cases like Pereboom's Case 1 and Case 2. Grant that the audience of neutral inquirers both would and ought rationally respond to Case 1 and Case 2 by ratcheting down their credence toward (M) in favor of an incompatibilist conclusion. And grant that this same audience both would and ought rationally respond to cases like my case of Ann and the Arpaly cases by ratcheting up their credence toward (M) and thereby move way from an incompatibilist conclusion. Suppose, furthermore, just for argument's sake, that the net overall effect for this audience both would and ought rationally be a lowering of credence toward (M) in favor of incompatibilism. On its face, this would seem to favor Pereboom's conclusion. But should we as philosophers just read off directly from these results our philosophical conclusions? I don't think so. I think there are good *theoretical* reasons to favor at least somewhat responses from neutral inquirers to cases like my case of Ann and the Arpaly cases as in contrast with cases like Pereboom's Case 1 and Case 2—and these theoretical reasons are *not* reasons to which we should expected an untutored audience to be especially sensitive.

This is a point I mentioned for a slightly different purpose in my original reply to Pereboom (McKenna, 2008: 157). It is reasonable to suppose that our intuitions have evolved along with our ordinary practices. Here, what I mean by intuitions are judgments in response to concrete cases expressing how pertinent concepts apply to them. These judgments, as Pereboom and I are using them, are meant to help settle the dispute between us. It is my contention that the further away from ordinary contexts the application of these concepts are, the less reliable we should take them to be. There's a natural explanation for this. Our training for these concepts

involved applications to contexts structured by our natural surroundings and our entire form of life. The more we move away from these, the less mooring we theorists have to feel confident that ordinary users are indeed applying our concepts properly. This is not to say that the application of our concepts in very outlandish contexts should be assigned no significance. It's just to say that *we as philosophers* have good reason to assign a higher degree of credence to those untutored judgments closer to ordinary contexts—*regardless* of the degree of confidence untutored subjects might assign to their judgments about outlandish cases when prompted. Cases like the case of Plum 1, for instance, seem so alien, and it can seem very easy to apply very confidently the concept of moral responsibility to the case. But as we have not evolved in a context in which it might be expected that some or all of us are caused to act as we do by teams of neuroscientists, rather than, say, our own DNA and our local environments, it is quite understandable that such contexts would generate a strong negative reaction. However, it is just not at all clear what our concept of responsibility would look like, or for that matter if we could even speak in terms of the same concept, if instead it evolved in the context of that very different sort of life.[21] Hence, as I remarked earlier (2008: 157), if incompatibilists such as Pereboom wish to barter dialectically in the currency of intuitions that are lifted out of the contexts that are their natural homes, then when they generate intuitions from wildly deviant cases, it is incumbent upon them to hold themselves to very high standards indeed.

In my opinion, once we secure these standards, compatibilists ought to concede to Pereboom that, for a neutral inquiring audience, case like his Case 1 and Case 2 place *some* degree-of-credence-lowering pressure on (M). Furthermore, it can even be granted that this credence-lowering pressure has the net effect of moving neutral inquirers to some degree in the direction of incompatibilism when counterbalanced with the credence-raising pressure of cases like my case of Ann or the Arpaly cases. But in light of the theoretical worries raised here, this should not be regarded as decisive. Even if, amongst neutral inquirers, intuitions about cases like Case 1 and Case 2 weigh more heavily than intuitions about my case of Ann and the Arpaly cases, they do not weigh heavily enough for the incompatibilists to claim victory. Maybe (M) is weakened, but it is not refuted.

Given the three preceding points, I conclude that Pereboom has not offered sufficient reason to believe that amongst an audience of neutral inquirers, his clarifying cases Case 1 and Case 2 give them sufficient reason to contend that (M) is false. And to the extent that his clarifying considerations give this audience good reason to lower somewhat their credence toward (M), there are theoretical reasons not to regard this as decisive in settling the dispute between compatibilists and incompatibilists. Admittedly, given my strategy of transferring the initially rational attitude it is appropriate to have toward Plum in Case 4, and given the concessions I have made here, (M) does not arrive at Case 1 untarnished. It's a bumpy ride, no

---

[21] Arpaly's lovely discussion of this point is especially insightful (2006: 109-13). In fact, as she makes the point, the mistake would be to have our responses to very unusual cases inform our responses to ordinary cases rather than the other way around. And, as she notes, in our actual practices, when some new, unusual event occurs that would place pressure on us to apply our old concepts to the new case, what we would do is seek anchor in our more common reactions and then carefully see how far we might apply them to the new case. What we would not do is take our reactions to the new case and then take this as occasion to revise our common, familiar, stable reaction to the more mundane cases.

doubt. But as a compatibilist pressing a hard-line reply, I claim that (M) does after all arrive at Case 1, even if it arrives weaker than it was when it first departed.[22]

9. Concluding Remarks

Many philosophers working on the issues of free will and moral responsibility will strenuously dispute my contention that the free will debate reduces largely to the manipulation argument. But amongst those source theorists who accept the soundness of Frankfurt's argument, or in some other way reject the premise that free will requires the ability to do otherwise, manipulation arguments are the only extant incompatibilist arguments withstanding careful examination. The ultimacy argument and the transfer of non-responsibility argument do not survive scrutiny. The former is question begging, and the latter relies for its credibility on something like the soundness of a manipulation argument. So, at least for source theorists, there is a great deal at stake in assessing manipulation arguments for incompatibilism.

Turning to manipulation arguments, I have made clear that there is no one-size-fits-all reply to the diverse range of them. Nevertheless, I have also argued that it is not a wise compatibilist strategy to shy away from a hard-line reply when doing so is a viable option. As regards the argument form, TMA, it is best to adopt a policy of inclining toward a hard-line reply. Soft-line replies to thoughtful instances are dialectically unstable. Careful philosophers like Pereboom are all too willing to revise their manipulation cases in ways that defeat a soft-line move. Eventually, we compatibilists have to face the music; some variation is going to turn out to be a tough one for us, and we're going to have to take a hard-line in response to it.

Derk Pereboom's four-case argument, perhaps alongside Alfred Mele's (2006) zygote argument, is the best version of a manipulation argument available. However, in adopting a hard-line reply to it, I have argued that, in the end, the extreme cases Pereboom relies upon, his Case 1 and Case 2, while damaging to the compatibilist thesis, are not decisive. A hard-liner ought to remain committed to claiming that these cases, properly fleshed out, are still such that the rational response to them by those whose intuitions ought to count is that it is not clear that the featured agents in them are not free and are not morally responsible.

In advancing my reply to Pereboom, several points have emerged that should be cause to reconsider the way the dialectic has unfolded thus far. One is that, so far as dialectical burdens are concerned, the compatibilist need only call into question the claim that an agent manipulated in the pertinent way is not free and is not morally responsible; she is under no obligation to argue that agents so manipulated *are* free and *are* morally responsible. Another is that manipulation cases far from ordinary contexts should be regarded with a much greater degree of skepticism than they often are when assessing their dialectical force. If so, this also suggests that the artillery the source incompatibilists have available is reduced even further, since it seems that they ought to restrict their arguments to cases that are closer to real life cases.

The preceding considerations suggest yet another interesting result. Incompatibilists might now consider exploring more fully currently available cases that are closer to real-life, or instead actual cases, that have the relevant features of manipulation cases and are (potentially) friendly to an incompatibilist diagnosis. Think, for instance, of Gary Watson's (1987)

---

[22] This shows how I am inclined to reply to an especially impressive incompatibilist move recently made by Patrick Todd (2011). Todd regards it as conclusive in favor of incompatibilism that intuitive responses to manipulation cases mitigate blame. Here, I am willing to make the concession regarding mitigation but resist the claim that it is decisive. (Space does not permit a proper rely to Todd here, although I intend to develop one elsewhere.)

exploration of the actual case of Robert Harris, whose ruthless murder of two innocent teenage boys can arguably be explained, at least to a great extent, by the extraordinary suffering and abuse Harris suffered throughout his entire childhood. Watson has suggested, and others have agreed, that a plausible explanation of one's ambivalence to hold Harris morally responsible for his crimes is an incompatibilist one. Or instead consider Mele's case of the goddess Diana, who created Ernie at a determined world just so he could one day perform a certain act. As Mele sets up the case, Ernie's history is no different from the history of any person created in the normal way, save for the first moments of his zygote's manufacturing process. While there is something very atypical about Ernie's beginnings and the intended purpose of his designer, Ernie's life is otherwise no different from the lives of many actual persons about whose behavior we have relatively reliable moral responsibility intuitions. The problem for incompatibilists, however, is that these cases do not seem to be so compelling in their effect of eliciting an intuitive response of non-responsibility. In fact, their force is so much less compelling that it seems the compatibilist is on dialectically safer grounds in arguing that in these cases the agents *do* act freely and *are* morally responsible. Thus, the compatibilist in these contexts need not so cautiously claim merely that it is not clear that the featured agents are not free and are not morally responsible. In these cases she can go the extra mile.

With all that has been said, doesn't it count *against* compatibilism that an attitude such as (M) is weakened by manipulation cases like Pereboom's? If such cases ought to move fair and open-minded neutral inquirers even somewhat in the direction of incompatibilism, doesn't this amount to a reason speaking in favor of incompatibilism? It seems to me that the honest answer is 'yes'; the fact that a case like Pereboom's Case 1 ought to move neutral inquirers even somewhat in the direction of incompatibilism is *a* reason in favor of incompatibilism. It's just that it's not a decisive reason.[23] Perhaps this will be regarded by some compatibilists as granting far too much. Some compatibilists (e.g., Dennett, 1984) tend to write as if there is not anything in the least counter-intuitive about compatibilism—that intuition unqualifiedly favors the compatibilist position. I disagree. Compatibilism is not an intuitive thesis through and through. It is an odd view that does after all bump up against some of our intuitions. Anyone's experience teaching the free will problem in an introductory philosophy class confirms the point. It is only to be expected that at points compatibilism winds up looking a bit embarrassed when exposed in nothing but its underwear. Cases like Case 1, I say, show compatibilism for what it is quite vividly—just as, Pereboom contends, Spinoza would have it. But it is a mistake to draw a conclusion against compatibilism for this sort of reason. Incompatibilism exposed in nothing but its tidy whities is no prettier; it too leads to counter-intuitive results. So, in the end, we need to weigh costs.

It is a hard truth that, as Harry Frankfurt remarked (2002: 28), we are inevitably fashioned and sustained by conditions over which we have no control. Compatibilists see it as a truth we can live with while persisting in regarding ourselves and others as free and responsible—even in the basic desert-entailing sense. Incompatibilists like Pereboom do not. Manipulation cases like Pereboom's Case 1 and Case 2 make it harder for the compatibilist to stomach that hard truth, but I have argued that they do not provide decisive grounds for giving up the compatibilist position altogether. They may very well lead us to pine for an existence in which we do fashion and sustain ourselves, at least to some extent, by conditions over which we

---

[23] My remarks here share affinities with Fischer's (2011) probing treatment of Mele's (2006) zygote argument. See especially Fischer (2011: 271-2).

do ultimately have control, so that there is no way that we could be anything like Plum in Case 1. But it is unclear whether or not this notion of being an ultimate creator and sustainer of oneself is even a coherent metaphysical possibility. Even if it is a metaphysical possibility, but one beyond the reach of limited creatures like ourselves, why should we conclude that, falling shy of it, nothing short of that is good enough?

Michael McKenna
University of Arizona

Bibliography

Arpaly, Nomy. 2006. *Merit, Meaning, and Human Bondage*. Princeton, NJ: Princeton University Press.

_____. 2003. *Unprincipled Virtue*. New York: Oxford.

Ayer, A. J. 1954. "Freedom and Necessity." In his *Philosophical Essays*. New York: St. Martin's Press: 3-20.

Buss, Sarah, and Lee Overton. 2002. *Contours of Agency: Essays on Themes from Harry Frankfurt*. Cambridge, Mass: MIT Press.

Campbell, Joseph. 1997. "A Compatibilist Theory of Alternative Possibilities." *Philosophical Studies* 88: 319-30.

Chisholm, Roderick. 1964. "Human Freedom and the Self." *The Lindley Lectures*. Copyright by the Department of Philosophy, University of Kansas.

Davidson, Donald. 1973. "Freedom to Act." In Honderich, ed., 1973: 67-86.

Demetriou, Kristin. 2010. The soft-line solution to Pereboom's four-case argument. *Australasian Journal of Philosophy* 88 (4): 595-617.

Dennett, Daniel. 1984. *Elbow Room: Varieties of Free Will Worth Wanting*. Cambridge, Mass: MIT Press.

Fara, Michael. 2008. "Masked Abilities and Compatibilism." *Mind* 117 (468): 843-65.

Fischer, John Martin. 2011. "The Zygote Argument Remixed." *Analysis* 71: 267-72.

_____. 2004. "Responsibility and Manipulation." *Journal of Ethics* 8 (2): 145-77.

_____. 1994. *The Metaphysics of Free Will*. Oxford: Blackwell Publishers.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.

Frankfurt, Harry. 2002b. "Reply to John Martin Fischer." In Buss and Overton, eds., 2002: 27-31.

_____. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68: 5-20.

_____. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 829-39.

Ginet, Carl. 1966. "Might We Have No Choice?" In Lehrer, 1966: 87-104.

Haas, Daniel. Unpublished. "In Defense of Hard-line Replies to the Multiple-Case Manipulation Argument."

Haji, Ishtiyaque. 2009. *Incompatibilism's Allure*. Peterborough, Ontario: Broadview Press.

_____. 1998. *Moral Appraisability*. New York: Oxford University Press.

Hobart, R. E. 1934. "Free Will as Involving Indeterminism and Inconceivable Without It." *Mind* 43: 1-27.

Hobbes, Thomas. 1651. *Leviathan*. R.E. Flatman & D. Johnston, eds. 1997. New York: W.W. Norton & Co.

Honderich, Ted., ed. 1973. *Essays on Freedom and Action*. London: Routledge & Kegan Paul.

Howard-Snyder, Daniel, and Jeff Jordan, eds., 1996. *Faith, Freedom, and Rationality*. Lanham, MD: Rowman and Littlefield.

Hume, David. 1748. *Enquiry Concerning Human Understanding*, ed. L.A. Selby-Bigge, 3rd ed. Revised, P.H. Nidditch, Oxford: Clarendon Press, 1975.

Hunt, David P.. 2000. "Moral Responsibility and Unavoidable Action." *Philosophical Studies* 97: 195-227.

Kane, Robert. 1996. *The Significance of Free Will*. Oxford: Oxford University Press.

Kearns, Stephen. 2011. Aborting the zygote argument. *Philosophical Studies*. Online April.

Lehrer, K. 1980. Preferences, conditionals and freedom. In P. van Inwagen (Ed.), *Time and cause: essays presented to Richard Taylor* (pp. 187-201). Dordrecht: D. Reidel.

_____, ed., 1966. *Freedom and Determinism*. New York: Random House.

_____.1968. "'Can's Without 'If's." *Analysis* 24: 159-60.

Lewis, David. 1981. "Are We Free to Break the Laws?" *Theoria* 47: 113-21.

Locke, Don. 1975. "Three Concepts of Free Action I." *Proceedings of Aristotelian Society* supp. Vol. IL: 95-112.

McKenna, 2011 (forthcoming). *Conversation and Responsibility*. Oxford University Press.

_____. 2008. "A Hard-line Reply to Pereboom's Four-case Argument" *Philosophy and Phenomenological Research* 77.1: 142-59.

_____. 2008b. "Frankfurt's Argument against Alternative Possibilities: Looking Beyond the Examples," *Nous* 42: 770-93.

_____. 2008c. "Saying Goodbye to the Direct Argument the Right Way." *Philosophical Review* 117.3: 349-383.

_____. 2008d. "Ultimacy & Sweet Jane." In Trakakis and Cohen, eds., 2008: 186-208.

_____. 2004. "The Relationship between Autonomous and Morally Responsible Agency." In *Personal Autonomy*. Ed., James Stacey Taylor. New York: Cambridge University Press.

_____. 2003. "Robustness, Control, and the Demand for Morally Significant Alternatives." In *Moral Responsibility and Alternative Possibilities*. Eds., Widerker and McKenna. 2002.

Mele, Alfred. 2008. Manipulation, compatibilism, and moral responsibility. *Journal of Ethics* 12, Nos.304: 263-86.

_____. 2006. *Free Will and Luck*. New York: Oxford University Press.

_____. 2005. "A Critique of Pereboom's Four-Case Argument for Incompatibilism." *Analysis* 65: 75-80.

Mele, Alfred, and David Robb. 1998. "Rescuing Frankfurt-Style Cases." *Philosophical Review* 107: 97-112.

Moore, G.E., 1912. *Ethics*. Oxford: Oxford University Press.

Pereboom, Derk. 2008. "A Hard-line Reply to the Multiple-Case Manipulation Argument." *Philosophy and Phenomenological Research* 77, No.1: 160-70.

_____. 2005. "Defending Hard Incompatibilism." *Midwest Studies in Philosophy* 29: 228-47.

_____. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

_____. 1995. "Determinism al Dente." *Noûs* 29: 21-45.

Schlick, Moritz. 1939. "When is a Man Responsible?" In Schlick, *Problems of Ethics* Prentice-Hall: 143-56. .

Schoeman, Ferdinand, ed. 1987. *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press.

Slote, Michael. 1982. "Selective Necessity and the Free-Will Problem." *Journal of Philosophy* 79: 5-24.

Smart, J.J.C. 1963. "Free Will, Praise, and Blame." *Mind* 70: 291-306.

Smilansky, Saul. 2000. *Free Will and Illusion*. Clarendon Press: Oxford.

Smith, Michael. 2003. "Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion." In Stroud and Tappolet, eds., 2003: 17-38.

Spinoza, Baruch. 1677/1985. *The Collected Works of Spinoza*, ed. and tr. E. Curley, Princeton: Princeton University Press.

Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Clarendon Press.

Stroud, Sarah and Christine Tappolet, eds., 2003. *Weakness of Will and Practical Irrationality* New York: Oxford University Press.

Stump, Eleonore. 1996. "Libertarian Freedom and the Principle of Alternative Possibilities." In Howard-Snyder and Jordan, 1996: 73-88.

Taylor, Richard. 1963. *Metaphysics*. Englewood Cliffs: Prentice Hall.

Todd, Patrick. 2011. "A New Approach to Manipulation Arguments. *Philosophical Studies* 152: 127-33.

Trakakis, Nick and Daniel Cohen, eds., 2008. *Essays on Free Will and Moral Responsibility*. Cambridge Scholars Publishing.

van Inwagen, Peter. 2008. "How to Think about the Problem of Free Will." *Journal of Ethics* 12, Nos.3-4: 327-41.

_____. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

_____. 1975. "The Incompatibility of Free Will and Determinism." *Philosophical Studies.* 27:185-99.

Vihvelin, Kadri. 2004. "Free Will Demystified: A Dispositional Account." In *Philosophical Topics* 32: 427-50.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA.: Harvard University Press.

Watson, Gary. 1987. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In Schoeman, 1987: 256-86.

Widerker, David. 2006. "Libertarianism and the Philosophical Significance of Frankfurt Scenarios." *Journal of Philosophy* 103: 163-87.

Widerker, David, and Michael McKenna, eds. 2003. *Moral Responsibility and Alternative Possibilities*. Aldershot, UK: Ashgate Press.

Wiggins, David. 1973. "Towards a Reasonable Libertarianism." In Honderich 1973: 31-62.

Wisdom, John. 1934. *Problems of Mind and Matter*. Cambridge University Press.

Zagzebski, Linda. 2000. "Does Libertarian Freedom Require Alternate Possibilities?" *Philosophical Perspectives* 14: 231-48.