Article

Theme: Author Meets Critics

Review of *Free Will, Agency, and Meaning in Life*, by Derk Pereboom

John Martin Fischer
University of California, Riverside
Email: john.fischer@ucr.edu

Mark Twain once said about Wagner's music: "It is not as bad as it sounds." Similarly, Derk Pereboom has convinced most of us working in this field that moral responsibility skepticism, at least of his kind, is not as bad as it sounds. Indeed, through his important work culminating (thus far) in this significant book, he has convinced me that the view is subtle, very plausible, and even deeply attractive. Whereas there are other forms of moral responsibility skepticism on the market, Pereboom's is (in my view) the most fully developed and most appealing. Although I am not myself a moral responsibility skeptic, and, in particular, I do not believe that causal determinism is incompatible with the kind of freedom that grounds robust, desert-based moral responsibility, I find that (surprisingly) our views don't really differ that much from each other. It might seem that our views are miles apart; but upon reflection, it turns out that they are very close, and with the intelligent and patient help of Pereboom, one can see the great appeal of his approach to a whole range of interrelated issues. I'm not going to mince words: this book is a masterful and comprehensive articulation of Derk Pereboom's very important and original theory of free will and moral responsibility. Throughout his career, and especially here in this book, Pereboom has developed and defended one of the real "contenders" as a comprehensive theory of freedom and responsibility. This is a huge, and admirable, intellectual achievement.

But Pereboom's contribution has not just been intellectual (narrowly construed). His gentle, patient, and constructive spirit leaps off the pages; reading his work, one feels that one is in contact with a genuinely kind and good person who is simply interested in getting to the truth. This helps to elevate the discussion and keep our eyes on the ball, as it were. He is open to criticism, reads widely and conscientiously, interprets critics charitably, and responds with great honesty. Pereboom's gentle spirit and his willingness honestly to engage his critics—and to read others' work widely and to take it seriously—is a model that is noteworthy and, again, elevates the discussion profoundly.

Pereboom offers penetrating critiques of event-causal and agent-causal libertarianism, and also compatibilism; below I shall focus on his critique of compatibilism. But before I do so, I wish to pause to commend what I take to be the most exciting and attractive parts of this book (and Pereboom's overall approach). Pereboom contends that the sort of moral responsibility at issue in the free will debate involves "basic desert". More specifically, he writes, "For an agent to be morally responsible for an action [in the relevant sense] is for it to be hers in such a way that she would deserve [in this basic way] to be blamed if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary." (2) He argues that it would be impossible for us to be morally responsible in this sense in a causally deterministic world, and that it is highly unlikely that we would be morally responsible in a causally indeterministic world.

But where does this leave us? This is where Pereboom's theory gets really interesting. Whereas other moral responsibility skeptics (such as, famously, the psychologist B.F. Skinner and the philosopher J.J.C Smart) argue that we need to give up our "responsibility practices" entirely and move to a system of positive and negative reinforcements (of various kinds), Pereboom presents a remarkably nuanced and sensi-

Smith & Franklin Academic Publishing Corporation

tive account of what we could maintain in our practices, even assuming moral responsibility skepticism. It turns out we could keep just about everything—or least everything we should, upon reflection, wish to keep. For example, we could simply prune our responsibility practices to eliminate resentment, hatred, and indignation; and many have argued that these are in any case unattractive elements of these practices.

Pereboom gives extraordinarily subtle and highly appealing accounts of the core of our responsibility practices (including "blame" and even "fury") that can be embraced by a moral responsibility skeptic. Further, he argues that a similar point applies to punishment and the criminal law: we can retain what we really should value, even on the assumption of moral responsibility skepticism. And so much the worse for the rest. Since it is indisputable that our criminal justice system is horribly flawed, Pereboom's alternative picture is attractive. Finally, he gives plausible accounts of practical reasoning, deliberation, and the meaning of life that dispense with desert-based moral responsibility and the associated kind of free will. Of particular interest is his insightful defense of a non-voluntaristic conception of love. Pereboom's claims may sometimes seem bolder than they really are; and yet they are sufficiently bold to be very interesting, while at the same time being defensible in a way that other bold ideas are not.

I particularly like Pereboom's discussion of love. It might have been thought that "true" or "genuine" love requires robust free will, if not in the "falling in love" part, at least in the "staying in love" part. Recall that in the film, "Bruce Almighty", God (played by Morgan Freeman) gave Jim Carey 's character special powers. This character was trying to woo the character played by Jennifer Aniston. At a crucial point, Morgan Freeman (playing God) tells Jim Carey's character, "You can do anything you want. But just don't mess with her free will." That seems to capture the voluntaristic conception of love; it suggests that genuine love requires freely accepting one's lover. Pereboom chips away at this view by starting with parental love for children (or children for parents), which is manifestly deep but apparently not based on free will at all. He extends the point to other forms of love, even romantic love. After all, it is not clear that either falling in love, or (upon reflection) maintaining the love, is a matter of freedom of the will. (For scholarly corroboration of the sort appropriate to the topic, see Elvis Presley, "Can't Help Falling in Love," and Frank Sinatra, "It Had to be You.") Pereboom's patient and nuanced discussion of these topics, especially as they fit into issues about the meaning of life, is a real high point in this book (and in Pereboom's work overall).

But Pereboom and I do have a significant disagreement about whether causal determinism is compatible with robust, desert-based moral responsibility. (Henceforth, I shall simply use "moral responsibility" [and related terms] to refer to this sort of moral responsibility—implying the notion that the individual in question deserves—in a "basic" way—reactive attitudes, such as moral praise and blame.) Pereboom contends that the strongest "anti-compatibilist" argument is the "manipulation argument" (really, in my view, a family of arguments, loosely united under the banner, "manipulation argument"). These arguments typically start with a scenario involving manipulation—an individual is "set up" in advance or directly manipulated by another agent who intends that the individual behave in a certain way. The proponent of the argument seeks here to elicit the intuition that the individual would not be morally responsible for his or her behavior, in such a scenario. This is the "no responsibility premise". Then the argument proceeds via a "no-difference" premise": a claim to the effect that there is no relevant difference between the manipulation scenario (in which there is no moral responsibility) and the "normal" situation under causal determinism. The conclusion then is that causal determinism is incompatible with moral responsibility.

This sort of argument has been around for a long time. Indeed, in my work I have taken worries about manipulation as very significant. I began my book, *The Metaphysics of Free Will: An Essay on Control* (Blackwell), with a set of thought-experiments (including the infamous "nefarious neurosurgeon") involving manipulation. (These thought-experiments caused Daniel Dennett, a philosopher not so inclined to take manipulation worries seriously, to implore, in his *Elbow Room: The Varieties of Free Will Worth Wanting* (MIT Press): "Please don't feed the bugbears!") I then asked the question: How do these thought-experiments differ from the typical or "normal" situation under causal determinism? My strategy was as follows: I agreed that causal determinism would share with these thought-experiments a crucial feature: ruling out freedom in the sense that requires alternative possibilities (freedom to do otherwise, or what

I dubbed, "regulative control"). But I further argued that the manipulation thought-experiments differed in a significant way from the ordinary situation under causal determinism: whereas manipulation (of the sort under consideration) expunges guidance control, causal determination *per se* need not. Thus, causal determination *per se* is compatible with moral responsibility, even granting that it (like the manipulation thought experiments) rules out regulative control.

Pereboom is not convinced, and he offers what is perhaps the most sophisticated and important version of a manipulation argument in contemporary philosophy: The Four-Case Argument. Anyone who wishes to defend compatiblism simply must address Pereboom's Four Case Argument. And whereas many of us (including Mele, McKenna, and I) have done so, Pereboom offers thoughtful and challenging replies in this book. Here I shall take the liberty of seeking (primarily) to develop and defend my critique further.

In the précis of the book provided by Gregg Caruso as part of this symposium, he summarizes the Four-Case Argument, and thus I shall simply provide the bare bones of the argument here. In all four of the cases, Professor Plum decides to (and does) murder White for some personal advantage. Pereboom claims that in all four cases the action under consideration, Plum's decision to kill White, meets the relevant compatibilist-friendly conditions for acting freely. And yet Pereboom claims that it is both clear that Plum is not morally responsible in the first case and that there is no relevant difference between the first case and the others. Thus, he contends, Plum is not morally responsible in any of the cases, including an "ordinary" case in which causal determinism obtains.

In case 1, "…neuroscientists manipulate Plum in a way that directly affects him at the neural level, but with the result that his mental states and actions feature the psychological regularities and counterfactual dependencies characteristic of genuine agency." (76) Following a suggestion of Seth Shabo, Pereboom here supposes that Plum uses an "egoism-enhancing mechanism" that momentarily heightens Plum's natural tendency toward egoism (but preserves his agency). The intuition about such a case—a case of "hands-on manipulation"—is surely that Plum is not morally responsible for his decision (and action). In case 2 neuroscientists have programmed Plum at the beginning of his life so that he is often but not always egoistic (as in case 1). When Plum decides to kill White and does so, Pereboom thinks that he is no more morally responsible than in case 1: he challenges a critic to point to a relevant difference. Similarly with cases 3 and 4. In case 3, Plum is otherwise an ordinary human being but the "training practices of his community causally determined the nature of his deliberative reasoning processes so that they are frequently but not exclusively rationally egoistic [as in cases 1 and 2]." (78) In case 4 Plum is supposed to be a normal human being (frequently but not exclusively egoistic) in a causally deterministic world. The idea of the argument is that Plum is clearly not morally responsible in the first case, and there is no relevant difference between the first case and the last.

This argument is challenging, and I have various different inclinations about how to respond. I probably have not been entirely consistent in my various published responses to manipulation arguments; and perhaps this reflects my uncertainty as to how exactly best to respond. But I *do* think that there are various promising responses to manipulation arguments. I certainly do not think they are decisive or so compelling that one must give up otherwise well-motivated theories (I shall return to this point below).

First, one might think that it is not obvious that Plum is not morally responsible even in the first case. This is because it is not clear how to evaluate the nature of the neural state that is induced by the team of neuroscientists. Pereboom writes about this neural state that it "realizes a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White." (76) But we do not yet really know what a "strongly" egoistic reasoning process is; we do not, for instance, know whether any reasoning process realized in this *kind* of way *must* issue in the decision to murder (and the act of murder). That is, we do not know whether a "strongly egoistic reasoning process" is *irresistibly* egoistic, in the sense that any process of *that type* would (and must) issue in egoistic choices and behavior. And this is true, even if one accepts the Consequence Argument or another argument to the effect that causal determinism implies that no one can choose or do other than he actually does.

To elaborate. If the neuroscientists induce a strong but not overwhelming or irresistible urge to murder White, then it is not obvious to me that Plum is not

morally responsible; he may, for all that has been said, be acting freely in this context. And this may be so, even if White could not have avoided his decision and act of killing White. (Here the notion of "irresistibility" is not simply "cannot be resisted in the particular circumstances". More, of course, would need to be said fully more adequately to articulate this "compatibilist" notion of irresistibility.) Now of course Plum's blameworthiness would be significantly diminished in this sort of situation; but, in my view, it would *not* follow that Plum is not morally responsible for his behavior. He is, after all, arguably at least an apt candidate for the reactive attitudes, even if we would not wish actually to target him with such an attitude in this situation (or would wish to target him with an attitude whose "degree" is reduced).

I think that the response just developed is plausible. But Pereboom disagrees, arguing that it is impossible for an agent to act wrongly and be morally responsible for the act in question but not blameworthy for it (89). I don't think Pereboom is obviously right here. One thing to note is that Pereboom relies heavily here on a conceptual point about the relationship between moral responsibility and blameworthiness about which (I would contend) reasonable people could disagree. But I shall proceed by simply stipulating, for the sake of discussion, that Pereboom is correct and that Plum is not morally responsible in case 1. It is, I admit, unclear what to say about this issue, and it is thus worthwhile exploring what my approach to moral responsibility can say about the four-case argument, on the assumption that Plum is not morally responsible in case 1.

In thinking about the cases, although these are extremely difficult and unclear issues, I am (at least sometimes) inclined to think that there is a crucial difference between case 1 and case 2; that is, arguably Plum is not morally responsible in case 1, but is in case 2. And my approach to moral responsibility is flexible enough to accommodate and explain this intuition. On my approach (together with Mark Ravizza), guidance control is the freedom-relevant condition for moral responsibility. And, importantly, guidance control has *two* components: reasons-responsiveness of the actual-sequence mechanism, and *ownership* of that mechanism. So, on our approach, an agent exhibits guidance control of his behavior insofar as it issues from his own, suitably reasons-responsive mechanism. Pereboom tends to focus solely on the reasons-responsive mechanism component, rather than the ownership component. But they are both indispensable, and the ownership part plays a crucial role in a proper analysis of manipulation cases.

It is salient, I think, that case 1 involves "hands-on", direct manipulation, whereas case 2 does not; case 1 is a genuine case of manipulation, whereas case 2 is perhaps better described as a case of "initial design". Although I laid out some reasons to think that Plum is morally responsible for his behavior in case 1 above, reasonable people can certainly disagree. Insofar as one has the intuition that Plum is not morally responsible in case 1, I believe that this is based on the view that the mechanism that issues in Plum's behavior is *not his own*. The agent's own mechanism of practical reasoning has been supplanted, and in its place we have a different kind of mechanism, inculcated surreptitiously by the neuroscientists. In contrast, even though the neuroscientists "set up" Plum with a set of initial dispositions in case 2, they do not subsequently intervene in a direct way, superceding his own mechanisms of practical reasoning. He has taken responsibility for his mechanism of practical reasoning against a backdrop of his "given" initial endowments. (This contrasts with case 1, in which Plum has taken responsibility for his mechanism of practical reasoning, but *not* for the "manipulation mechanism" inculcated by the neuroscientists.) The situation in case 2 is not relevantly different from the ordinary situation in which we are simply "given" a set of dispositions toward feeling and action; and moral responsibility always then is a matter of how one plays the cards that are dealt one, as it were. I think there is an important difference between a case of direct, hands-on manipulation, such as case 1, and a case of initial design, such as case 2; and I contend that this difference lies in the fact that Plum acts from his own mechanism in case 2 but not in case 1.

When discussing case 2, Pereboom writes (in a footnote): "Fischer and Ravizza (1998) hoped to preclude manipulation by specifying that moral responsibility requires an agent to come to understand and accept that she is morally responsible through a reflective historical process that involves rational sensitivity to the evidence. But since this historical reflective endorsement is a causal process, the neuroscientists can manipulate Plum to realize it…" (87) Pereboom here does not appear to credit our explicit recognition that the beliefs that are part of mechanism ownership can

indeed be manipulatively implanted or our insertion of an *additional* condition intended to rule it out that we would deem such a mechanism the agent's own.

Just so that we have the Fischer/Ravizza account of mechanism ownership on the table, note first that we claim that a kind of mechanism becomes the agent's own through a historical process of "taking responsibility". We start with two belief conditions on taking responsibility, and then we add a third condition constraining the etiology of the beliefs specified in the first two conditions: "First, an individual must see himself as the source of his behavior… That is, the individual must see … that his choices and actions are efficacious in the world." (Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*: 210) "Second, the individual must accept that he is a fair target of the reactive attitudes as a result of how he exercises this agency in certain contexts". (Fischer and Ravizza, 211) "The third condition on taking responsibility requires that the individual's view of himself specified in the first two conditions be based, in an appropriate way, on the evidence." (Fischer and Ravizza, 213)

We explicitly observe that the first two conditions can be met in a context of manipulative inculcation of the relevant beliefs (Fischer and Ravizza, 235-6), and we sketch a way in which the third condition can address this worry (especially in footnote 31 on 236-7). This is not the venue in which to debate whether our suggestion works; I admit that it needs considerable further development. But it is not fair for Pereboom to reject our approach so precipitously, as if we did not consider the problem of manipulative inculcation of "taking responsibility". And it is puzzling that Pereboom would contend that we would seek to avoid pinning responsibility on Plum in case 2; indeed, I am inclined to employ the resources of taking responsibility and mechanism ownership to drive a wedge between cases 1 and 2. As above: arguably, at least, Plum acts from his own mechanism in case 2, but not case 1. On this sort of approach, Plum would be morally responsible for his decision and action in case 2, but not case 1.

My suggestion, then, is that one might invoke guidance control to distinguish case 1 from the rest of the cases in Pereboom's Four-Case Argument. This, of course, was the strategy of my overall argument in *The Metaphysics of Free Will* (referred to above); I there claimed that in thought-experiments involving direct manipulation, guidance control is absent. But I argued that this point cannot be extrapolated to the "ordinary" case, under causal determinism. Similarly, I here argue that what is problematic about case 1 is the lack of guidance control; but, again, this point cannot be extrapolated to the rest of the cases.

Step back from the details and think of it this way. A compatibilist wants to say that not all causally deterministic sequences are relevantly similar in threatening free will/moral responsibility. A semicompatibilist can (although he need not) say that all such sequences equally rule out freedom to do otherwise; in this respect they are just like manipulation scenarios involving nefarious neurosurgeons. But it would not follow that all causally deterministic sequences equally rule out "actual-sequence freedom" and moral responsibility. I hold this view, and I have a suggestion for what makes the difference between the different kinds of causally deterministic sequences; the factor I invoke is guidance control. Other compatibilists invoke other factors in seeking to make this distinction; for example, Mele and others invoke the notion of "bypassing". I think that even if none of the suggestions on the table is fully acceptable (and I certainly recognize that reasonable people can be unconvinced, especially because there are important gaps in my development of guidance control), nevertheless one can still feel the pull of the claim that there is *some* relevant difference, even if it is hard to specify. And this would seem to be enough to defend compatibilism (or, at least, semicompatibilism). After all, there are places in philosophy where one feels strongly that there is *some* difference between two sets of phenomena, even if it is hard to characterize the difference in a fully adequate, reductive way. In this sort of situation, it is sometimes legitimate to stick with the distinction, even if one cannot explain it fully.

But when? I think this raises complex and delicate dialectical issues. I often think that these dialectical issues are crucial in evaluating the debates about manipulation cases (and the associated arguments)—and yet they are frequently under-appreciated. Here's one way to think about the situation. (I fully recognize that it is highly contentious, and it is clear that Pereboom does not share this view about the dialectical situation: 80-82) Semicompatibilism is at least plausible and it offers many considerable attractions: robust moral responsibility and full personhood do not hang

on a thread (i.e., are not held hostage to the arcane deliverances of theoretical physicists), we can make distinctions between those who are free and morally responsible and those who are not (distinctions that line up with common sense and reflective theorizing on the basis of common sense), and we can avoid deep and intractable metaphysical disputes about the relationship between prior truths and human freedom, God's foreknowledge and human freedom, and causal determinism and human freedom. So there is at least a *prima facie* case for semicompatibilism.

Now the view faces an objection (or set of objections) based on manipulation, initial design, and related worries. So we semicompatibilists here need to "play defense". Let's say you play defense in the fourth quarter. You do not need to win the fourth quarter in order to win the game. Even if you were to lose the fourth quarter by a touchdown, if you go into it with a two touchdown lead, then you still win the game. And, in my view, that's how it is with regard to semicompatibilism and the manipulation argument; in playing defense here, it is enough to sketch various approaches on which we achieve a rough draw. So I concede that the challenges are real and difficult, but I also maintain that compatibilists (especially semicompatibilists) have various resources at their disposal with which to construct plausible defenses.

Note that Pereboom's view is that what rules out moral responsibility in *all* of his four cases is that the behavior is "produced by a deterministic process that traces back to factors beyond the agent's control." (73) He presents this view "upfront", but I'm not sure that this reflects a moment in dialectical history, as it were, rather than simply an accidental feature of his presentation. It should be obvious that we couldn't make any progress in *any* of these debates if one party were simply to *start with this assumption and stick to it no matter what*. That is, if one party to the disputes simply started with the intuition that whenever any behavior is deterministically caused by a process that traces back to factors beyond the agent's control, then the agent is not morally responsible, no progress could possibly be made. This is essentially simply to *start with* incompatibilism and not to let any reflection on potential differences among (putatively) different kinds of causally deterministic sequences affect one's views. That is a dialectical disaster waiting to happen.

Now I'm not saying that Pereboom employs this strategy, but it sometimes seems like something like this is lurking in the background of at least some discussions of the Four Case Argument (and related arguments). A better way to think of the dialectic, and probably the way Pereboom is indeed thinking, is that we at first suspend judgment. We do not *start* with the view that whenever any behavior is causally determined by a process that traces back to factors beyond the agent's control, then the agent is not morally responsible. Rather, one reflects carefully and with an open mind about the cases and then concludes that this is the *best explanation* of one's intuitions about the cases. But if this is the way to think about the dialectical context, then it makes salient the issue of what our considered intuitions should indeed be, and what precisely *is* the best explanation of those intuitions. It is far from clear to me that Plum is not morally responsible in all four cases. And it is equally unclear to me that the best explanation of his moral responsibility status is the straightforward incompatibilist view invoked by Pereboom. I think that many would be inclined to make subtler discriminations, especially upon careful reflection. As regards the issues of our *considered* judgments and their best explanation, I do not think that the Four Case Argument is an obvious victory for incompatibiism.

Pereboom writes:

> The four-case argument serves to draw attention to the deterministic causes of action that would be present if in general our actions were causally determined, but which nevertheless typically are hidden from us… In his response, Fischer [citation suppressed] contends that we can make a distinction between two kinds of hidden causes, the first of which impairs responsibility, the second not. The first kind interferes with the normal functioning of mechanisms, while the second 'is simply the set of constituents of the overt properties'—these are the more specific or concrete ways in which the overt properties are instantiated in the neural structure of the brain. Fischer would contend that if the brain is functioning properly, the neural instantiation of properly reasons-responsive deliberation and action will not threaten our intuitive judgments of moral responsibility, even if the neural structure were governed by deterministic laws.

> Fischer's key claim is that hidden causes of the second sort pose no threat to moral responsibility even if they are governed by deterministic laws. I disagree, and I base this judgment on my manipulation argument, and more generally, my case against compatibilism rests on the strength of this argument. (90-91).

Note again that Pereboom simply refers to "properly reasons-responsive mechanisms". But guidance control requires that these mechanisms be *the agent's own*. That is, it requires that the agent have taken responsibility for the relevant kind of mechanism. As I wrote above, it is plausible that Plum's mechanism is *not* his own in case 1, as opposed to the other cases. I do not see how Pereboom's four-case argument in itself, or supplemented by ancillary argumentation, provides a strong reason to resist this contention. And so I do not see how it provides a strong reason to deny that hidden causes of the second sort pose no threat to moral responsibility, even if they are governed by deterministic laws. It would only provide such a reason if it contained a clear example of an agent who acts from *his own*, suitable reasons-responsive mechanism in a deterministic scenario, and it is obvious that he is not morally responsible for his behavior. But I have argued that in the case in which it is perhaps closest to being obvious that the agent is not morally responsible, Plum is (arguably, at least) *not* acting from *his own*, suitably reasons-responsive mechanism.

I would insist, then, that nothing in Pereboom's argumentation licenses him to reject the distinction between the two kinds of hidden causes. He conceptualizes the point of the four-case argument as indicating that if we were to recognize that we are causally determined by hidden causes, we would give up our view of ourselves as free and morally responsible. I think it is at least equally plausible to suppose that the Four Case Argument has a different function: it points to the possibility that we are causally determined by a *special kind* of hidden cause—perhaps one that does not leave room for guidance control, or one that bypasses one's normative structures, or… In such a *special case*, one would not be morally responsible, even though one is not aware of the problematic hidden causes. But it would not be warranted to extrapolate to a general conclusion about *all* causally deterministic sequences.

In doing administrative work in my academic institution, I was taught that it is always best to offer a "Feedback Sandwich". So having given the meat of my critique, I return to some more fluffy and complimentary carbs. This book contains significant new developments and defenses of Pereboom's comprehensive theory, which embraces free will, moral responsibility, criminal justice, practical reasoning, and meaningfulness in life. Pereboom's views are remarkably nuanced and appealing. And I have just scratched the surface. In this review, I have not been conscientious in citing the big literature on these topics—a literature in part due to Pereboom's influential work. Notably, Pereboom does not share this vice; his engagement with critical work is admirable. This book will offer even more reason for philosophers to think seriously about a view that many of us have under-appreciated: moral responsibility skepticism. I guess it is not as bad as it sounds. In fact, it is pretty darn impressive.

Acknowledgements

This paper was written during the period of support from a grant from the John Templeton Foundation. I am very grateful for this support, but the views here expressed do not necessarily reflect those of the Templeton Foundation.