Philos Stud (2011) 152:127–133
DOI 10.1007/s11098-009-9465-8

A new approach to manipulation arguments

Patrick Todd

Published online: 23 October 2009

Abstract There are several argumentative strategies for advancing the thesis that moral responsibility is incompatible with causal determinism. One prominent such strategy is to argue that agents who meet compatibilist conditions for moral responsibility can nevertheless be subject to responsibility-undermining manipulation. In this paper, I argue that incompatibilists advancing manipulation arguments against compatibilism have been shouldering an unnecessarily heavy dialectical burden. Traditional manipulation arguments present cases in which manipulated agents meet all compatibilist conditions for moral responsibility, but are (allegedly) not responsible for their behavior. I argue, however, that incompatibilists can make do with the more modest (and harder to resist) claim that the manipulation in question is *mitigating* with respect to moral responsibility. The focus solely on whether a manipulated agent *is* or *is not* morally responsible has, I believe, masked the full force of manipulation-style arguments against compatibilism. Here, I aim to unveil their real power.

Keywords Free will · Moral responsibility · Manipulation · Derk Pereboom · Compatibilism · Incompatibilism

1 Introduction

There are several argumentative strategies for advancing the thesis that moral responsibility is incompatible with causal determinism. One prominent such strategy is to argue that agents who meet compatibilist conditions for moral responsibility can nevertheless be subject to responsibility-undermining

P. Todd (✉)
Department of Philosophy , The University of California, Riverside, HMNSS Building, Room 1604,
900 University Avenue, Riverside, CA 92521, USA
e-mail: Pat.c.todd@gmail.com

manipulation. In this paper, I argue that incompatibilists advancing manipulation arguments against compatibilism have been shouldering an unnecessarily heavy dialectical burden. Traditional manipulation arguments present cases in which manipulated agents meet all compatibilist conditions for moral responsibility, but are (allegedly) not responsible for their behavior. I argue, however, that incompatibilists can make do with the more modest (and harder to resist) claim that the manipulation in question is *mitigating* with respect to moral responsibility. The focus solely on whether a manipulated agent *is* or *is not* morally responsible has, I believe, masked the full force of manipulation-style arguments against compatibilism. Here, I aim to unveil their real power.

I present my case by investigating what is (so far) the most sophisticated manipulation argument against compatibilism: Derk Pereboom's widely influential "Four Case" argument. There is much to be appreciated in this argument. However, as I hope to show, it merely contains the seeds of a more powerful one.

2 Modifying the four case argument

Consider "Case 2" (perhaps the central case) of Pereboom's "Four Case" argument for incompatibilism:

> Plum is like an ordinary human being, except that a team of neuroscientists has programmed him at the beginning of his life to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the consequence that in the circumstance in which he now finds himself, he is causally determined to undertake the reasons-responsive process of deliberation and to possess the set of first- and second-order desires that result in his killing White. Plum does have the general ability to regulate his behavior by moral reasons, but in his circumstances the egoistic reasons weigh heavily for him, and as a result he is causally determined to murder White. Nevertheless, he does not act because of an irresistible desire[1] (Fischer et al. 2007, p. 75).

Is Plum morally responsible (in the sense that he deserves blame) for killing White? Pereboom says 'no'. However, and this is the central point, the case is designed so that Plum fully meets all compatibilist conditions for moral responsibility. The upshot, then, is that compatibilist conditions for responsibility are too

[1] Case 2 has been interpreted in various different ways, some ways being 'stronger' than others. In particular, Mele (2005) reads the case in such a way that Plum is importantly different with respect to control than 'typical' agents. However, as I read the case (and as I will understand it in this paper), it is no different than what Mele has called an 'original design' scenario. Mele considers the case of a goddess Diana who creates a zygote in an environment such that, with the laws of nature, it is determined that the resulting person (Ernie) will do X after 30 years (2006, p. 188). In other words, Ernie is like everyone else, except that the details of his life were the result of the intentional activity of Diana. Mele considers this case to be severely problematic for the compatibilist. I regard Case 2 and Mele's 'Zygote Argument' scenario to be on a par. The neuroscientists simply 'set up' Plum in such a way that he is determined to kill White. Moreover, Pereboom has confirmed in conversation that he means Case 2 to be understood in this way.

weak: one can meet such conditions, as Plum did, yet nevertheless be subject to the responsibility-undermining manipulation of the neuroscientists.

Of course, one possible compatibilist reply to this case would have it that there is some principled distinction between one's acts being determined by the neuroscientists and one's acts being determined by mere natural causes. And, so this strategy says, all the compatibilist is committed to is the thesis that moral responsibility is compatible with mere causal determinism, *not* that it is compatible with determinism by manipulators. Thus, there must be some (or we must provide some) further compatibilist condition for responsibility that is in fact violated in Case 2, but which would *not* be violated if mere determinism holds.

But Pereboom argues (convincingly, I think) that there is no relevant difference between Case 2 and mere causal determinism. Pereboom compares Case 2 to *another* case, Case 4, where Case 4 is simply Case 2 over again, except that natural deterministic causes have taken the place of the neuroscientists. By comparing such cases, Pereboom says, we can see that it is in fact *irrelevant* that Plum's act was deterministically brought about by such manipulators; whether Plum's psychological states ultimately trace back to intentional agents or non-intentional causes should not matter. We can thus generalize from Plum's lack of responsibility in Case 2 to Plum's lack of responsibility in Case 4. Moreover, it is worth noticing that leading compatibilists have in fact *agreed* that there is no principled distinction between one's acts being determined by the neuroscientists and by mere natural causes. Hence, the rest of this paper will simply take for granted the thesis that there is no relevant difference vis-à-vis blameworthiness between Case 2 and Case 4. Of course, compatibilists are free to try to articulate such a difference. Here I simply note that those taking such a line face a difficult challenge, and that I do not believe this strategy is promising.

Now, notice: Pereboom's argument asks the reader to concur in the judgment that the victim of Case 2-style manipulation is simply *not responsible*—not at all. To begin to see how Pereboom's argument can be modified, I suggest that we re-imagine the case along the following lines. Suppose one is an eye-witness to White's horrific murder, but one does not yet know anything about the role of the neuroscientists. The murder seems to be (in the relevant respects) 'typical'—one can see that Plum was not coerced into performing the act, not acting on any compulsive desire, that he murdered White for selfish reasons, knew what he was doing, and so on. Now, we imagine that the following question is put to one:

Q1 On a scale from 1 to 10, rate how much blame Plum deserves for killing White, where 0 is no blame at all, and 10 is the most blameworthy you can imagine someone being.

Now, one writes down one's answer to this question. Of course, the question could be different in some respects; the question could ask, for instance, how many years in prison (or some such) Plum deserves (on non-consequentialist grounds) for his act. The important point here is that one writes down one's judgment about Plum *before* one finds out about the neuroscientists.

So we now imagine that the broader picture is unveiled: one sees how the neuroscientists programmed Plum in such a way as to make his killing White causally determined. Now, manipulation arguments so far have (implicitly) gone on to ask the question (and *solely* the question):

Q2 Having now found out about the role the neuroscientists played in programming Plum, do you still think Plum deserves blame for killing White?

The incompatibilist response to Q2 is of course to say: *no*. The incompatibilist thinks that, having found out about the neuroscientists, one should no longer think that Plum deserves blame for killing White. As the incompatibilist sees things, after having found out about the neuroscientists, one should judge that if *anyone* deserves blame for Plum's killing White, it is the *neuroscientists* and not Plum.

Now, the compatibilist answer to Q2 is of course to say: *yes*. That is, the compatibilist maintains that despite having found out about the neuroscientists, it is still appropriate to judge that Plum deserves blame for his act. The central dispute concerning manipulation arguments has thus far simply been whether a *yes* or *no* response to Q2 is the appropriate one. As I'll now argue, sole focus on this question is a mistake.

To begin to see why focusing solely on a yes/no response to Q2 is a mistake, recall the question originally put in Q1:

Q1 On a scale from 1 to 10, rate how much blame Plum deserves for killing White, where 0 is no blame at all, and 10 is the most blameworthy you can imagine someone being.

Again, the compatibilist believes that, even being aware of the role of the neuroscientists, it remains appropriate to think Plum is blameworthy, or to think Plum still deserves punishment, and so on. But, of course, this position is consistent with radically *revising* one's initial judgment of blameworthiness after gaining full information. That is, one could answer *yes* to Q2, but with the following qualification: "I initially rated the amount of blame Plum deserves as a 7 out of 10. Now, I still feel that he deserves blame, but with full information, I think he deserves only a 4 out of 10, rather than a 7 out of 10. I don't feel nearly as badly towards him."

In other words, while the incompatibilist straightaway judges that any '7' she might have initially felt towards Plum should now be reduced to a '0', others may not go all the way here—they may hold that while the manipulation in question *diminishes* Plum's blameworthiness, it does not *eliminate* it. But what would such a judgment mean for compatibilism? Seemingly, it would mean that the truth of determinism implies *diminished* or *mitigated* blameworthiness. Recall: there is (very plausibly) no important difference between Plum's acts being brought about by the neuroscientists and by merely natural but deterministic causes. So if one judges that the role of the neuroscientists in Plum's life (setting him up the way he is, etc.) implies lessened blameworthiness, one should likewise judge that the role of impersonal deterministic causes implies lessened blameworthiness.

So, unless she admits that determinism implies mitigated blame, the compatibilist is in fact committed to something much stronger than a mere 'yes' to Q2; she is committed to the claim that finding out about the role of the neuroscientists should make *no difference* to one's feelings of moral disgust towards Plum. The compatibilist seemingly must endorse what we might call the

No Difference Thesis: Case 2-style manipulation should make *no difference* to one's judgment of how much blame Plum deserves for killing White.

As the incompatibilist sees things, it is hard enough to maintain that Plum deserves blame, but things are apparently even harder: if the compatibilist is to be believed, not even a *revision* of judgment is appropriate, given knowledge of Plum's background. I (along with other incompatibilists) submit that this is an excessively strong claim.

Now, perhaps some compatibilists are prepared to agree with me that the No Difference Thesis is excessively strong. That is, perhaps some compatibilists have looked into their (heretofore hardened) philosophical hearts and seen that they *would* (or should) feel less inclined to harshly judge (or punish) Plum, given the lousy lot he received at the hands of the neuroscientists. I welcome the softness of these hearts. But, they argue, all compatibilism *qua* compatibilism is committed to is, well, the thesis that blameworthiness is *compatible* with determinism, not that the truth of determinism is simply *irrelevant* to blameworthiness. Hence, they maintain, compatibilism remains undefeated, despite an admission of lessened blameworthiness if determinism is true. If the compatibilist takes this line, a new challenge presents itself.

Here is the new challenge. Again, the compatibilist in question is someone who *admits* that the truth of determinism implies mitigated blameworthiness. But if the compatibilist admits that determinism itself is mitigating, a fair question is, In virtue of what? What is it about determinism's obtaining that makes revised judgments of blameworthiness appropriate? Here the compatibilist is on thin ice, for she must specify features of determinism that only *mitigate* responsibility rather than *ruling it out*. Now, what could such features be? I submit that I cannot see what the compatibilist could offer here. For instance, the compatibilist may say: determinism mitigates responsibility because its truth would entail that the characters from which our choices flow are partially the result of factors beyond our control. This is right, of course, but it is of course also right that if determinism is true, the characters from which our choices flow are *entirely* the result of factors beyond our control—there is apparently no room for degrees here. And if the fact that our characters are partially the result of such factors is sufficient to diminish responsibility, surely the fact that they are *entirely* the result of such factors is sufficient to eliminate it.

Or perhaps the compatibilist says: the truth of determinism implies mitigated blameworthiness because determinism rules out alternative possibilities. (Notice: no one could plausibly maintain that determinism merely rules out *some* alternatives.) But surely this is awkward. Presumably, if alternative possibilities have any role to play in moral responsibility, it is that they are *necessary* for it, not just that they are an inessential 'add-on' which merely deepens or increases moral responsibility, and without which it remains (basically) intact.[2] At any rate, any view on which

[2] Indeed, those who accept that the upshot of so-called 'Frankfurt-style cases' is that alternative possibilities are not necessary for moral responsibility typically think that these examples show that alternatives are *irrelevant* or *useless*. That is, if the examples work as supposed, they appear to show that *nothing* important depends on alternatives, since (apparently) everything important to agency is nevertheless retained by Jones (the monitored agent that, due to the presence of the counterfactual intervener, has no alternatives). Jones is *no less blameworthy*, despite the fact that he could not do otherwise.

alternative possibilities play this sort of role would need to be articulated and defended.

So, we are left wondering: according to the compatibilist that rejects the No Difference Thesis, why does determinism mitigate the amount of blame our bad actions call for, or the amount of punishment they deserve? I do not see any obvious compatibilist reply to this question. Of course, I do not claim to have *proven* that compatibilists cannot consistently reject the No Difference Thesis, but minimally I believe the dialectical burden would be on the compatibilist to articulate a plausible picture of what such a rejection would look like. In the absence of such a picture, any compatibilist rejection of the No Difference Thesis must appear strained.[3]

I believe we are now in position to briefly state how a manipulation argument with the new approach I favor may be articulated. We may call this argument the Modified Manipulation Argument (here on the MMA):

(1) If blameworthiness is mitigated for Plum in Case 2, blameworthiness is mitigated if mere causal determinism is true.
(2) If blameworthiness is mitigated if mere causal determinism is true, then compatibilism is false.
(3) Blameworthiness is mitigated for Plum in Case 2.

So, (4) Compatibilism is false.

Now, I have just been defending (2)—the claim that compatibilists cannot plausibly maintain that determinism merely *mitigates* blameworthiness. Further, I believe (1) to be relatively uncontroversial. And here, with other incompatibilists, I simply assert my considered judgment that (3) is true: Plum's is a case of mitigated blame. Given the premises, the conclusion that compatibilism is false follows.

But perhaps the compatibilist will disagree with my considered judgment about Case 2, despite its being weaker than a judgment that Plum is simply *not* blameworthy. That is, one might think that the best compatibilist reply to the MMA is *not* to reject (2), but to stick to one's guns, and maintain that one's initial reaction to Plum should not weaken after becoming aware of the neuroscientists. Once it is understood, compatibilists may say, that Plum is not subject to compulsive desires, is fully able to regulate his behavior by moral reasons, is fully reasons-responsive, and so on, we will have no reason to weaken our negative attitudes towards him. Thus, one denies (3) and maintains the No Difference Thesis: Case 2 style manipulation is not mitigating for Plum.

But here it is worth drawing some conclusions. If denying (3) is the best compatibilist response to the MMA, then incompatibilists wielding manipulation arguments have been shouldering an unnecessarily heavy dialectical burden. All incompatibilists should be (or must be) claiming is that Case 2-style manipulation (or its equivalent) *dampens* or *detracts* from Plum's blameworthiness: a judgment of

[3] While I am skeptical that a plausible compatibilist rejection of the No Difference Thesis will be forthcoming, we ought to see that such a rejection would still be a significant admission on the part of compatibilists. For, in reading most compatibilist literature, one does not get the sense that compatibilists believe that the truth of determinism *in any way* threatens the correctness or the appropriateness of our judgments about moral desert. A rejection of the No Difference Thesis would be an admission that this position has been wrong: determinism *is* relevant.

'7' should at least decrease to a '6', and so on, *not* that a 7 should decrease to a *zero*. This burden is *significantly* lighter than the one incompatibilists have so far been carrying, and the burden is to that extent significantly heavier for compatibilists. Is it really plausible to think that the fact that Plum got such a raw deal at the hands of the neuroscientists is simply irrelevant to Plum's moral desert? I do not think so, but such a result appears to be the (increased) cost of compatibilism.

3 Conclusion

In this paper, I have argued that traditional manipulation arguments against compatibilism can be modified so as to be much stronger than heretofore supposed. Incompatibilists need only the judgment that the relevant kind of manipulation (whether it be Case 2-style or otherwise) *mitigates* blameworthiness. If such arguments—as is widely held—already have shown that compatibilists must take a hard line, I hope to have shown that this line is yet harder still.[4]

References

Fischer, J., Kane, R., Pereboom, D., & Vargas, M. (2007). *Four views on free will*. Malden, MA: Blackwell.

Mele, A. (2005). A critique of Pereboom's 'Four Case Argument' for incompatibilism. *Analysis, 65*(285), 75–80.

Mele, A. (2006). *Free will and luck*. Oxford: Oxford University Press.

[4] For helpful comments on previous versions of this paper, I wish to thank my colleagues and fellow members of the agency reading group at the University of California, Riverside, Justin Coates, Chris Franklin, Benjamin Mitchell-Yellin, Garrett Pendergraft, Philip Swenson, and especially Neal Tognazzini. Kevin Timpe and Michael Nelson also provided valuable comments on an earlier draft. I am deeply indebted to John Martin Fischer for valuable comments on several drafts of the paper. Lastly, for many helpful conversations about these topics, I wish to thank those who participated in Central European University's July 2009 'Aspects of Responsibility' summer program, where a version of this paper was presented. In particular, I wish to thank Michael McKenna, Paul Russell, and Derk Pereboom.