Philosophical Studies (2023) 180:487–507
https://doi.org/10.1007/s11098-022-01906-2

Manipulation, machine induction, and bypassing

Gabriel De Marco[1]

Accepted: 21 November 2022 / Published online: 8 December 2022

Abstract
A common style of argument in the literature on free will and moral responsibility is the Manipulation Argument. These tend to begin with a case of an agent in a deterministic universe who is manipulated, say, via brain surgery, into performing some action. Intuitively, this agent is not responsible for that action. Yet, since there is no relevant difference, with respect to whether an agent is responsible, between the manipulated agent and a typical agent in a deterministic universe, responsibility is not compatible with the truth of determinism. In response, some theorists have argued that there is a relevant difference, and have developed two sorts of accounts of that difference: bypassing views, and manipulator-focused views. Manipulator-focused views suggest that the difference concerns the presence of a manipulator, whereas bypassing views suggest that the relevant difference concerns the fact that the action issues from attitudes that the manipulated agent acquired in a way that bypassed her capacities for control over her mental life. One sort of case used to decide between these sorts of accounts is a case of machine induction, which is just like a manipulation case, yet the change in the agent is the result of some natural force. Against the received view, Xiaofei Liu argues that such cases pose problems for bypassing views, and favor manipulator-focused views instead. This paper addresses Liu's arguments, as well as a variety of cases, concluding that cases of machine induction do not provide motivation for a bypassing theorist to adopt a manipulator-focused view.

Keywords Moral responsibility · Manipulation argument · Machine induction · Bypassing

✉ Gabriel De Marco
gabriel.demarco@philosophy.ox.ac.uk

[1] Faculty of Philosophy, Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK

1 Introduction

One prominent way to argue for the claim that determinism is not compatible with moral responsibility is to offer a manipulation argument. Such arguments tend to involve cases of agents in deterministic universes who, due to substantial manipulation they are not aware of—e.g., brain surgery to modify values—perform some action. Though such agents are stipulated to meet standard conditions on responsibility for action—e.g., they have control at, or shortly before, the time of action—they do not seem to be responsible for the action. According to manipulation arguments, there is no relevant difference, with respect to whether agents are responsible, between such manipulated agents and typical agents in deterministic universes; call this the *no-difference thesis*. Thus, such arguments conclude, typical agents in deterministic universes are not responsible for their actions, and determinism is not compatible with moral responsibility.

The two main sorts of replies to manipulation arguments are so-called soft-line and hard-line replies.[1] Hard-line replies adopt the counter-intuitive position that the manipulated agent is responsible for his action, while accepting the no-difference thesis. Soft-line replies accept the intuitive claim that the manipulated agent lacks responsibility for his action, and deny the no-difference thesis. This sort of reply is accompanied by an account of the relevant difference.[2]

We can roughly divide soft-line responses into two camps, one which we can call *bypassing views*, and another which we can call *manipulator-focused views*. Manipulator-focused views suggest that, in order to explain why subjects of this sort of manipulation are not responsible for the relevant actions, we will need to appeal to the presence of a manipulator.[3] Since typical agents are not manipulated in these ways, there is a relevant difference between them and the sorts of agents we find in manipulation arguments. Alternatively, on bypassing views, part of the difference between the manipulated agent and a typical agent in a deterministic universe is that the attitudes leading to the manipulated agent's action were acquired or modified in a way that did not engage, and thus bypassed, his capacities for control over his mental life—e.g., the capacities to assess, endorse, and sustain or modify one's values in light of this reflection.[4] From now on, I will simply refer to this process as *bypassing*.[5]

One sort of case used to distinguish between the two sorts of soft-line reply are cases that do not involve manipulators. Such cases are very much like manipulation

[1] This distinction, as it is commonly used, comes from (McKenna, 2008).

[2] Not all responses fit neatly into this division. For instance, see (Haji & Cuypers, 2001; Kearns, 2012; King, 2013; Takasaki, 2021).

[3] These views differ in detail, but for examples, see (Barnes, 2015; Deery & Nahmias, 2017; Herdova, 2021; Usher, 2020; Waller, 2014; Yaffe, 2003). One notable exception is Deery and Nahmias's view, which can apply to some cases in which there is no manipulator (Deery & Nahmias, 2017, pp. 1272–3).

[4] For slightly different capacities, or fuller lists of these capacities, see (Fischer, 2012, p. 198; Haji & Cuypers 2008, p. 30; McKenna, 2016, p. 97; Mele, 1995, pp. 118–120, 2019, p. 45).

[5] For examples of bypassing views, see (De Marco, 2021; Fischer, 2012; Fischer & Ravizza, 1998; Haji & Cuypers, 2008; McKenna, 2016; Mele, 2019).

cases, yet the change in the agent occurs as a result of some force that is not an agent.[6] Consider a version of such a case recently offered by Xiaofei Liu, based on a case presented by Derk Pereboom in his famous four-case manipulation argument (Pereboom, 2014, Chapter 4, 2017):

> *Machine Induction*. Plum grows up in an environment that is saturated with radio signals randomly and spontaneously sent out by a machine, which is not designed or controlled by any agent. These signals work directly on Plum's sensory organs to cause him not only to have the character, the reasoning patterns and the value system that he has, but also to think in particular ways in various circumstances by, for instance, presenting certain stimuli to arouse particular reactions. Despite all this, the signals happen to work in such a way that Plum thinks just like an ordinary person and he satisfies all the compatibilist conditions for moral responsibility (for example, he is moderately reasons-responsive and his effective first-order desires conform to his second-order volitions). Under the causal influence of these radio signals, Plum designs a plan and kills White. (Liu, 2022, p. 536)[7]

This sort of case is often taken to pose a problem for manipulator-focused views. Since there is no manipulating agent, we cannot explain Plum's lack of responsibility for killing White by appealing to a manipulator.

In his recent paper, Liu pushes back on this line of reasoning, and argues for two main claims. First, he argues that this case does not pose a problem for manipulator-focused views, since there is no relevant difference between Plum in *Machine Induction* and a typical agent in a deterministic universe. Once we recognize this, we ought to think that Plum is responsible for killing White. Second, he argues that bypassing views face a dilemma when confronted with *Machine Induction*. For a particular bypassing view: "if it denies responsibility in *Machine Induction*, it will be forced to reject compatibilism; if, on the other hand, it affirms responsibility in *Machine Induction*, it will be forced into a hard-line position" (Liu, 2022, p. 546). The main goal of this paper is to push back on this dilemma. Yet in doing so, I also show that, depending on how we understand *Machine Induction*, there may be reason to reject the no-difference thesis with respect to Plum and a typical agent in a deterministic universe. Thus, what I suggest will also provide grounds for rejecting Liu's first conclusion. If there is a relevant difference between typical agents in deterministic universes and Plum in some versions of *Machine Induction*, and we think that Plum is intuitively not responsible in these variations, then some versions of *Machine Induction* remain a problem for manipulator-focused views. Before

[6] See, for example, (Mele, 2019; Mickelson, 2019; Pereboom, 2014).

[7] Liu also presents a second case, *Multi-Machine Induction* (Liu, 2022, p. 537). The difference between the two cases does not concern the nature, frequency, or randomness of the signals, only whether the signals originate in one machine or many (Liu, 2022, p. 538). Though this difference may (or may not) be relevant for how we apply Deery and Nahmias's view (Liu, 2022, pp. 539–43), the two cases can be treated as the same for present purposes. Thus, I simply focus on *Machine Induction*.

doing this, however, it is important to get clear about some features of *Machine Induction*, and some assumptions I will be making.

One feature of the case is that the radio signals cause Plum to have the character, reasoning patterns, and value system that he has, and to think in particular ways in various circumstances. Yet, as others have pointed out in response to previous versions of Pereboom's cases, in order for Plum to be anything like a typical agent in a deterministic universe, these signals cannot be the only causes of these features of Plum (Baker, 2006; Demetriou, 2010; Vihvelin, 2013, p. 152).[8] For instance, for a process to count as reasoning, or an instance of deliberation, some of the mental states involved need to be causally related to some of the other mental states in appropriate ways.[9] As I will be understanding the case of *Machine Induction*, then, Plum's mental states have similar causes as those of typical agents in deterministic universes, including his own previous mental states. That is, Plum's mental states at one moment will tend to be among the causes of the mental states that he has in the next moment. Yet on top of this, there is also this machine which spontaneously sends out random signals which influence Plum's mental states, in some way or other. Thus, when it comes to Plum's character, reasoning patterns, etc., there are many other causal influences on them—e.g., other mental states—which one might expect to find in a deterministic universe which does not have this machine.[10]

Now consider a different feature that the case of *Machine Induction* might have: the signals are present, and have this influence, throughout Plum's entire life. Alternatively, these signals might have an influence on Plum only occasionally. The more that these signals influence Plum's character, reasoning patterns, and value system—and the more instances of such influence—the more complicated the case will get. In order to home in on the details of the case, and how bypassing views might apply, I will begin by understanding the case as a simple one in which this is the first time that the radio signals have an influence on Plum's mental states. With this discussion in place, I return to variations of the case.

This paper proceeds as follows. First, I introduce bypassing views in more detail, with a focus on Liu's main target: Mele's bypassing view. Then, I consider each horn of the dilemma, and argue that upon closer examination of *Machine Induction*, the dilemma fails. Finally, I sketch potential responses to further variations of the case in Sect. 4.

[8] For somewhat related points, see also (Fischer, 2006, Chapter 12; Mele, 2005).

[9] Alternatively, this might need to be true of the neural correlates of these mental states.

[10] Matheson (2016) offers case which may avoid these worries, and Pereboom recently adopted this version (Pereboom, 2017). However, even if one grants that Matheson's case serves this purpose, these worries still leave us with significant restrictions on what a manipulation case must be like in order to avoid them. This is enough, for current purposes.

2 A primer on bypassing views

According to bypassing views, the sorts of subjects we tend to find in manipulation cases are not responsible for the relevant action, in part, because their actions issue from attitudes acquired via bypassing. However, this is not all that there is to it. As many have pointed out, we are subject to a variety of influences in our daily lives, and it is quite possible that many of these influence our attitudes via bypassing.[11] Thus, insofar as one is attempting to offer a response to cases of manipulated agents that avoids the counter-intuitive conclusion that we are not responsible for much of what we do, one should reject the claim that an action's issuing from an attitude that was acquired via bypassing is sufficient to eliminate an agent's responsibility for that action. [12] And bypassing theorists do seem to reject this.[13]

Further, there are cases of much more moderate manipulation in which it is intuitive that the agent's responsibility for a particular action has not been undermined. Consider, for example, Mele's case of Carl, who has made a commitment to refrain from eating snacks for 6 months, yet experiences, every day, a few medium-strength desires to eat a snack. Although the urge is always resistible, he occasionally acts on it. Suppose now that a manipulator induces in Carl such an urge about once a day, and Carl succumbs to it about 5% of the time. As Mele suggests, Carl's "being morally responsible for eating snacks in response to such urges is implausibly regarded as turning on whether the urges are produced, on the one hand, in the 'normal' way or…, on the other, by a manipulator who flashes subliminal 'snack' messages at him" (Mele, 2019, p. 37).

In order to avoid these problems, bypassing theorists offer more nuanced accounts. Since Liu's main target is Mele's view, we can focus on that one here. As Liu characterizes the view, "a necessary condition for an agent to be autonomous is that those pro-attitudes that produce the action are possessed authentically" (Liu, 2022, p. 543). And, in order for a pro-attitude to be authentic, the agent must not be compelled to possess it (Mele, 1995, p. 166).[14]

There is, however, an important clarification to make. The view described here is of autonomous *possession* of an attitude (Mele, 1995, p. 156). Yet, this does not imply that a necessary condition on autonomous *action* is that the pro-attitudes producing the action are authentic. In later work, Mele expands his view to freedom and responsibility with respect to actions, and adds further conditions (Mele, 2006,

[11] See, for instance, (Arpaly, 2006; Fischer, 2012; Frankfurt, 2002; McKenna, 2017; Mele, 1995).

[12] One might worry that this begs the question against some theorists—e.g., skeptics about responsibility—who deny that we are responsible for much, if any, of what we do. For the purposes of this paper, I assume that we are responsible for much of what we do. The main discussion of this paper, as well as Liu's argument (as I read it) involves an in-house dispute among compatibilists, and perhaps more broadly, those who think that ordinary agents are responsible for many of their actions.

[13] For some discussion, see (Fischer, 2012, pp. 196–200; Haji, 1998, p. 132, 2010, p. 278; McKenna, 2017, pp. 579–80; Mele, 2019, pp. 35–8, 54–5, 130–3).

[14] Mele distinguishes between "compulsion" and "compulsion*." The latter is compulsion that the agent did not arrange for (Mele, 1995, p. 166). I follow Liu in replacing "compulsion*" with "compulsion."

Chapter 7, 2019). Mele offers the following necessary condition on direct responsibility for an action:

DMR. If an agent is directly morally responsible[15] for *A*-ing, then the following is false:

(1) for years … his system of values[16] was such as to preclude his acquiring even a desire to perform an action of type *A*, much less an intention to perform an action of that type;
(2) he was morally responsible for having a long-standing system of values with that property;
(3) by means of very recent [bypassing] to which he did not consent and for which he is not morally responsible, his system of values was suddenly and radically transformed in such a way as to render *A*-ing attractive to him during *t*; and
(4) the transformation ensures either,

a that although he is able during *t* intentionally to do otherwise than *A* during *t*, the only values that contribute to that ability are products of the very recent [bypassing] and are radically unlike any of his erased values (in content or in strength) or,
b that, owing to his new values, he has at least a Luther-style "inability" during *t* intentionally to do otherwise than *A* during *t*. (Mele, 2019, pp. 127–8)[17]

Conjunct 4 mentionsLuther-style inability, in reference to Dennett's discussionof the phrase famously attributed to Martin Luther: "Here I stand, I can do noother" (Mele, 2019, pp. 62–4). The most concisecharacterization is expressed by Dennett when he states that: "when I say Icannot do otherwise I mean I cannot because I see so clearly what the situationis and because my rational control faculty is *not* impaired" (Dennett, 1984, p. 133). Notably, this senseof (in)ability is concerned with doing otherwise in relevantly similarcircumstances.

One thing to notice is that, according to *DMR*, one might still be directly responsible for an action that issues from an attitude one is compelled to possess.If, for instance, one is able to do otherwise in the relevant sense, and there are values contributing to this ability that are not the result of the radical transformation one underwent via bypassing—as described inconjuncts 1 and 3—then 4 is false, and thus the conjunction of 1–4 is false.[18] Further, notice that an application of *DMR* to Carl, in

[15] To say that an agent is directly morally responsible for A-ing is, roughly, to say that she is responsible for A-ing, and that this responsibility does not wholly stem from responsibility for some further B (Mele, 2019, p. 11).

[16] I follow Mele (1995, p. 116, 2019, p. 14) and McKenna (2016, p. 88) in their understanding of "S values X": "S at least thinly values X at a time if and only if at that time S both has a positive motivational attitude toward X and believes X to be good."

[17] I have slightly modified the condition to remove references to a manipulator. As Mele makes clear elsewhere, his view is intended to apply to cases in which there is no manipulator present as well (Mele, 2019, pp. 27, 58).

[18] For discussion of a similar possibility, see (Mele, 2008, p. 269, n. 13).

the case of moderate manipulation described above, would not tell us that he lacks responsibility for eating the snack, when that results from the implanted attitude.

With this fuller picture of bypassing views, andMele's in particular, we can now turn to Liu's dilemma.

3 A return to the dilemma

Recall that Liu states that for any soft-line reply like Mele's—that is, for any bypassing view—either:

(1) It denies that Plum is responsible for killing White in *Machine Induction*, in which case it is forced to reject compatibilism, or
(2) It affirms that Plum is responsible for killing White in *Machine Induction*, in which case it is forced to take a hard-line approach.

When introducing each horn of the dilemma, Liu stipulates details of *Machine Induction*. These further details, I suggest, result in at least two different versions of the case. This dilemma fails, I argue, once we consider how the details of the case of *Machine Induction* are to be filled out, since the bypassing theorists have different responses available, depending on which version of the case we are considering. We can approach each horn of the dilemma individually.

3.1 The first horn

How might one argue for the first horn? As a first step,

> [s]uppose that we judge that Plum's possession of the relevant pro-attitudes in *Machine Induction* is compelled, because Plum's capacities for control over his mental life was bypassed, the bypassing issued in Plum's being practically unable to shed those pro-attitudes, the bypassing was not itself arranged by Plum, and so on. (Liu, 2022, p. 544)

Once we make this supposition, we can apply Mele's view to show that Plum does not act from authentic attitudes when he kills White. Liu takes this to imply that on Mele's view, Plum is thereby not responsible for the killing (Liu, 2022, p. 545).

Yet, as we saw above, the fact that an action issues from an attitude one is compelled to possess is not enough to undermine responsibility for that action. The fact that one is compelled to *possess* a certain attitude, for instance, is not sufficient for one's being compelled to *act* on that attitude. In order to get this horn of the dilemma started we need to establish the further claim that according to bypassing views—and Mele's in particular—Plum is not responsible for killing White. What we need to assume, then, is that Plum fails to meet a necessary condition on responsibility put forth by bypassing views. Since we are focusing on Mele's view, we can suppose

that Plum fails to meet *DMR*.[19] Thus, suppose that shortly before the machine sends out its signals, Plum's system of values was such as to preclude him from acquiring even a desire to kill White. The signals, which work via bypassing, suddenly and radically transformed his system of values such that killing White is now attractive to him, and the transformation ensured that either he has a Luther-style inability to do otherwise at the relevant time, or if he has this ability, it is only due to further values that were a result of the sudden and radical transformation. Call this version of the case, *Strong Machine Induction*, or *SMI* for short. With respect to *SMI*, Mele's view would tell us that Plum is not responsible for killing White.

How do we proceed from here to get the rest of the first horn of the dilemma? Liu appeals to a version of the no-difference thesis. Specifically, Liu suggests that "if we believe that Plum's control over his mental life is bypassed in *Machine Induction*, we must also believe that the same is true in the [case of ordinary causal determination]. That is to say, we would be forced to reject compatibilism" (Liu, 2022, p. 545).

How might the bypassing theorist respond? Notice that in this claim, Liu only appeals to the fact that Plum's capacities for control over his mental life were bypassed. Yet the bypassing theorist can accept that this happens to ordinary agents in deterministic universes without denying compatibilism, insofar as the fact that agents' capacities for control over their mental lives are bypassed isn't sufficient to undermine responsibility. Liu must have more than this in mind. Suppose we get more specific, and suggest that if we believe that Plum's action issues from attitudes he acquired via bypassing, we must also believe this of other actions in deterministic universes. If this is what is intended, then the bypassing theorist can offer a twofold response.

First, the bypassing theorist can deny the claim that believing that Plum's action issues from attitudes acquired via bypassing in *SMI* commits us to the claim that every action performed by an agent in a deterministic universe is also the result of attitudes acquired via bypassing. The truth of determinism does not make it such that agents do not have these capacities, nor does it make it such that agents never exercise them when acquiring new attitudes, nor does it make it such that agents only act from attitudes acquired via bypassing.

Second, the bypassing theorist can point out that even for those instances in which agents in deterministic universes act on attitudes acquired via bypassing, this still won't be sufficient to undermine responsibility for those actions. Recall the case of Carl, who acts on the basis of an attitude he acquired via bypassing, yet is still responsible for the action; both according to bypassing views, and plausibly, our intuitions. Thus, even for those instances in which ordinary agents in deterministic universes act on the basis of attitudes acquired via bypassing, these actions may still differ from Plum's killing of White in relevant ways. The agent may not be compelled to possess these attitudes, nor need it be the case that there was a radical

[19] Structurally, this is not significantly different than Liu's suggestion that we assume Plum is compelled to possess the relevant pro-attitudes, since he takes that to be why Plum would fail to meet Mele's necessary condition on responsibility for an action.

reversal of the sort described in *DMR*, nor need it imply that when an agent acts from such an attitude, they either have a Luther-style inability to do otherwise, or if they have the relevant ability, it is only due to other attitudes acquired in the radical reversal.[20]

Thus, if we interpret *Machine Induction* as *SMI*, the generalization from Plum's action in SMI to other actions in ordinary deterministic universes fails, and the dilemma is avoided; the bypassing theorist can appeal to her bypassing view to show a relevant difference. We may, however, flesh out the details of *Machine Induction* in a different way.

3.2 The second horn

On the second horn of the dilemma, the bypassing theorist is meant to face the problem that, if she affirms that Plum is responsible for killing White in *Machine Induction*, she is forced to take a hard-line approach; that is, she is ultimately forced to adopt the counter-intuitive position with respect to manipulated agents. In arguing for this horn, Liu proceeds in a similar fashion, by asking us to make some suppositions about the nature of the case.

According to Liu,

> one may insist that, in the case of ordinary causal determination, Plum's possession of the relevant pro-attitudes is not compelled–for example, his capacities for control over his mental life were not bypassed, or such bypassing did not issue in Plum's being practically unable to shed those pro-attitudes, or such bypassing was arranged by Plum himself, and so on. But since we cannot find a control-relevant difference between *Machine Induction*…and a case of ordinary causal determination, we would then have to say that Plum in *Machine Induction* is also not compelled and thus should be morally responsible for the killing. Thus, in order to save compatibilism, one would be forced to hold that Plum's possession of the relevant pro-attitudes in *Machine Induction* is not compelled. (Liu, 2022, p. 545)

On this route, we begin by considering a typical agent in a deterministic universe who is not compelled to possess the attitudes leading to his action—in fact, whose attitudes leading to action were not acquired via bypassing—and, on the basis of a no-difference thesis, come to a similar conclusion about Plum in *Machine Induction*. Could the bypassing theorist resist this move? As we saw above, a bypassing theorist could resist a generalization from *SMI* to a typical agent in a deterministic

[20] A reviewer helpfully points out that this horn of the dilemma could be understood as applying to a more restricted set of cases of agents in deterministic universes, cases introduced by Arpaly (Arpaly, 2002, p. 127, 2006, pp. 109–116) and further appealed to by others, such as McKenna (McKenna, 2008, pp. 156–7), Cyr (Cyr, 2020, p. 2390), and Shaw (Shaw, 2014, pp. 7–8). Notice, however, that if such cases were a problem for the bypassing theorist, this still does not justify the claim that she is committed to rejecting compatibilism, since it is just for a subset of cases. Further, bypassing theorists have responded to such cases, arguing that they do not pose a problem for their views (Haji & Cuypers, 2008, pp. 58–60; Mele, 2006, pp. 179–84, 2020, p. 3148).

universe. If we interpret *Machine Induction* as *SMI*, then, the bypassing theorist can resist the move from a typical agent in a deterministic universe to Plum in precisely the same way.

In order to make the move from a typical agent in a deterministic universe to Plum in *Machine Induction*, then, we will need a version of the case that does not involve an influence as radical as the one in *SMI*. What might such a case look like? Suppose we go with a version on which Plum, as with the typical agent in a deterministic universe Liu mentions, has not had his capacities for control over his mental life bypassed. Liu seems to provide a case that might be interpreted in such a way:

> The radio signals could, for example, simply constitute some situational cues that critically prompt [Plum] to think in a particular way about his situation, much like an ordinary actor being prompted to act by some provoking words of a friend who knows the actor well…If an ordinary actor can be morally responsible when her action is prompted (without loss of her compatibilist friendly agential structure) by some uninvited provoking words, why cannot [Plum] be responsible when his action is prompted in the same way by situational cues from some machines? (Liu, 2022, p. 539)

Let us further suppose that, like the typical agent in a deterministic universe, Plum is not compelled to possess these attitudes, the change induced by the radio signals does not amount to the radical reversal described in *DMR*, etc. Call this version of the case *Weak Machine Induction*, or *WMI* for short. If we interpret *Machine Induction* as *WMI*, then Liu is right to suggest that the bypassing theorist cannot point to a relevant difference between a typical agent in a deterministic universe and Plum in *WMI*.

How do we proceed from this to get the rest of the second horn of the dilemma? We can simply modify the case such that, rather than a machine, it is now a manipulator sending out the signals (Liu, 2022, p. 545). Call the Plum in this case *Weakly Manipulated Plum*. Just like the bypassing theorist cannot point to a relevant difference between a typical agent in a deterministic universe and Plum in *WMI*, she also cannot point to a relevant difference between Plum in *WMI* and Weakly Manipulated Plum. Thus, the bypassing theorist is forced to accept that Weakly Manipulated Plum is responsible for killing White.

How might a bypassing theorist respond? There are a few points to make in response. First, consider a similar modification of *SMI*, on which everything is the same, but for the fact that it is a manipulator sending the signal. Call this Plum *Strongly Manipulated Plum*. A bypassing theorist can still point to a relevant difference between Weakly Manipulated Plum and Strongly Manipulated Plum, and she is not committed to a hard-line reply to the latter.[21] After all, there are substantial differences between these two Plums outlined by a bypassing view; and as is often recognized, there is no "one-size-fits-all response" to such cases or arguments

[21] It may be worth noting here that much of the discussion around manipulation cases and bypassing views revolves around cases much more like that of Strongly Manipulated Plum than cases like that of Weakly Manipulated Plum. See, for instance, (Haji & Cuypers, 2008; McKenna, 2016; Mele, 2019).

(McKenna, 2008, p. 143; Mele, 2019, p. 118).[22] At most, this second horn of the dilemma forces the bypassing theorist to adopt a hard-line response to Weakly Manipulated Plum. This, one might think, is not *too* hard of a line to take.

But the bypassing theorist might push even further and suggest that this is not even a hard-line *at all*, insofar as the claim that Weakly Manipulated Plum is responsible for killing White is not counter-intuitive. One could point out that, since Plum did not acquire the relevant attitudes via bypassing, this means that the acquisition engaged Plum's capacities for control over his mental life; e.g., the capacities to assess, endorse, and sustain or modify one's values in light of this reflection. Taking this into account, one might think that the mere fact that the attitudes originated from a manipulator-operated machine will not be enough to undermine Plum's responsibility for killing White.

The bypassing theorist might further point out that, *even if* the signals influenced Plum's mental states via bypassing, yet the influence was fairly mild, and resulted in a small effect, we might still not get the intuition that Plum lacks responsibility. Recall, for instance, Mele's case of Carl above. Or consider McKenna's suggestion that

> If the cause introduced is no different in any relevant respect than the way that, for instance, a momentary alteration in attention due to bad digestion might affect someone's deliberation or subsequent decision, or a quick spike in blood sugar, or an unexpected remark about one's abusive father…and if all the other control elements are held in place and operate in non-deviant fashion, it is hard to see why we should think that an agent's freedom or control is impaired...and I suspect that neutral inquirers could be brought rather easily to revise their initial intuitive reactions to a case like Case 1 once it was shown to them how little intervention would be needed by this team of neuroscientists to achieve its desired effect. (McKenna, 2017, pp. 579–80)[23]

Thus, if we interpret *Machine Induction* as *WMI*, with or without bypassing, the bypassing theorist can still argue that accepting that Plum in *WMI*, and Weakly Manipulated Plum, are responsible for their killings does not commit them to a hard-line reply, insofar as this is not a counter-intuitive stance.

To sum up, Liu's argument for each horn of the dilemma involves making different suppositions about Plum in *Machine Induction*, suppositions which the bypassing theorist takes to be relevant for responsibility. Once we evaluate both horns while holding details of the case fixed, however, the bypassing theorist has responses available. If we interpret *Machine Induction* as *SMI*, the bypassing theorist can claim that Plum is not responsible for killing White, yet is not committed to a rejection of compatibilism. If we interpret it as *WMI* however, she can claim that Plum *is* responsible for killing White, as is Weakly Manipulated Plum, while

[22] For an extended discussion of this, and related, points, see (Sekatskaya, 2018).

[23] When McKenna makes this point, he is offering a development of his original hard-line reply to at least one version of the case. However, the quoted sentence is consistent with the sort of reply I am suggesting here, and coheres well with Mele's points concerning the case of Carl.

holding either that this hard-line response is not a very hard one, or push further and deny that it is a hard-line response at all.

4 Variations

With the main response in place, I can now briefly consider some potential variations of *Machine Induction*, and sketch some responses available to the bypassing theorist.

4.1 A lifetime of signals

One might worry that there is something problematic in assuming that, in *Machine Induction*, this is the first time that the signals affect Plum. Dropping this assumption, and considering a case in which Plum is receiving these signals fairly commonly, and throughout his life, may pose further problems for the bypassing theorist.[24]

Modifying *Machine Induction* in this way certainly does make things more complicated. First, describing such a case in detail would require much more than has previously been done in this literature; one cannot, briefly, flesh out the details of all of these influences that an agent has had over a lifetime.[25] We may, then, have to resort to a relatively high level of abstraction when describing such cases. Second, it becomes more difficult to determine what a bypassing view would say about such a case, insofar as there are now quite a few different individual influences to consider; applying the views to such a case will be more complicated. Third, it becomes more difficult to make judgments about these cases. Would we judge Plum to lack responsibility for killing White in such a case? It is not clear. As with the question of how bypassing views would apply, this might depend on the nature of the various different influences. And assessing such a case would require us to keep in mind various details about many influences—in this case, over an agent's lifetime—and this is not easy to do. If, instead, we describe it at a relatively high level of abstraction, it may be difficult to form a clear judgement. Thus, it may be more difficult to form the relevant judgments about such cases. Insofar as making comparative assessments of bypassing views and manipulator-focused views partly involves applying the views to particular cases and determining whether they align with our judgments about that case—that is, whether or not the views yield counter-intuitive results—then these sources of difficulty— in describing the case, applying the views to it, and

[24] It may be worth noting a further complication: if such a world is like what we might think of as an ordinary deterministic world, but *also* has these signals influencing agents, then such agents might generally have mitigated responsibility, and this is for reasons that even hard-liners to all manipulation cases could accept. For instance, if the signals are not reason-conferring, yet causally contribute to their actions, this might make such agents generally less sensitive to reasons (Kaiserman, 2021, p. 706).

[25] For an example of what I have in mind, see (Mele, 2019, pp. 19–21) for descriptions of two cases that involve just one intervention and one relevant action (or, perhaps instead, one short-term plan).

forming of judgments about it—can make it difficult to use such cases to assess the views.

Here, and without developing all of the intricacies of such a case, I briefly consider some variations where the nature of the influences is stipulated. First, suppose that Plum is constantly undergoing the sorts of changes he does in *SMI*. In every one of these changes, Plum goes from being such that his previous system of values would preclude certain actions that are now attractive to him, and either he has a Luther-style inability to do otherwise when it comes to these actions, or he has the relevant ability, but this is only due to further values that were a result of the sudden and radical transformation. Call such a case *A Lifetime of Strong Machine Induction* (*LSMI*). What problem does *LSMI* pose for the bypassing theorist? It does not change how they might account for the claim that Plum is not responsible for killing White. If Plum recently underwent an influence akin to that found in the single-influence *SMI*, he still fails to meet *DMR* with respect to his action of killing White.[26] Further, the bypassing theorist can account for Plum's lack of responsibility for various other actions that are the result of these radical reversals. If one thinks that Plum in *SMI* is not responsible for killing White, then adding further *SMI*-like influences in Plum's past does not seem to create a further problem for the bypassing theorist.

Second, suppose that Plum's attitudes are influenced only slightly by these signals, in the sort of way that we saw in *WMI*. Call such a case *A Lifetime of Weak Machine Induction* (*LWMI*). One might still worry that, since these signals are present throughout Plum's entire lifetime, Plum is, in some sense, being buffeted about by the machine's signals throughout his life. It is, however, important to keep in mind not just the fact that his attitudes are influenced by these signals, but also the nature of these influences. These signals, when they influence Plum, may slightly modify his values; but Plum is not compelled to possess the values he acquires, and they are not the result of a radical reversal. Further, when these attitudes issue in action, Plum still meets standard conditions on responsibility for actions; e.g., he has control at, or shortly before, the time of action. If the signals are like the manipulator's influence in the case of Carl, then they would also seem fairly easy to resist.[27] If, as in the first version of *WMI*, these influences do not even tend to bypass Plum's capacities for control over his mental life, and thus engage with them, then it is not clear in what way he would be buffeted about by these signals, nor is it clear that this makes a difference to whether he is responsible for killing White. But even if they *do* tend to bypass his capacities for control over his mental life, yet we hold other things fixed—e.g., he is not compelled to possess these new attitudes, he has control

[26] With respect to *DMR* in particular, one might worry that this is not the case, insofar as it is questionable whether conjunct 2—stating that he "was morally responsible for having a long-standing system of values with that property"—is true of Plum in this instance, given that he has had previous *SMI*-like influences. Whether it does might depend on various other features of the case and prior influences. However, Mele suggests elsewhere that excluding this conjunct from *DMR* would still result in a sufficient condition for lacking responsibility for *A*-ing (Mele, 2019, pp. 136–7).

[27] Mele stipulates that they are resistible, but also that he only succumbs to them about 5% of the time. This low rate would seem close to implying relatively easy resistibility.

at, or shortly before, the time of action, etc.—then it is still not clear that he would lack responsibility for killing White. The influences from these signals may not be relevantly different from influences we already encounter in our everyday lives via, for instance, various marketing techniques, product design, or influences like what Richard Thaler and Cass Sunstein call nudges (Thaler and Sunstein, 2021).[28]

Third, suppose that the influence prior to Plum's deciding to kill White is just like the one in *WMI*, yet Plum's past includes a variety of different signals. Might this make a difference to the bypassing theorist? This might depend on the nature of these earlier signals. Suppose, for instance, that the night before the events of *WMI* took place, Plum was subjected to signals akin to those found in *SMI*, and the signals in *WMI* just made a slight difference to how he deliberated from this new system of values that he acquired the day before. We might think that in such a case, Plum is not responsible for killing White. Yet, the bypassing theorist can account for this, since Plum in this case would fail to meet *DMR*.

4.2 A spectrum of influences

A second sort of variation of the case of *Machine Induction* does not involve stipulating that Plum receives signals throughout his lifetime, but rather involves varying the effects of those signals. To keep things simple, we can return to cases in which Plum is influenced by the machine only once, and shortly before killing White.

One might worry that the sorts of changes that Plum undergoes in *WMI* and *SMI* are extreme ends on a spectrum; either they involve a radical change, or a very mild one. But one can imagine that the signals result in a change in Plum's mental states more substantial than the one found in *WMI*, yet not as radical as the one in *SMI*. For instance, they might create stronger or weaker attitudes in Plum, or they might modify existing attitudes to a greater or lesser extent, both in terms of the number of attitudes changed, and the extent to which individual attitudes are changed. Further, the changed or acquired attitudes can be more or less central to Plum's system of values, and they can hold different positions in his hierarchy of values. And Plum may have more or less of an opportunity to evaluate the influenced attitudes, and more or less of a capacity to do otherwise than kill Plum.[29] Thus, although the bypassing theorist can avoid Liu's dilemma when we focus on *SMI* and *WMI*, things are not so simple when we consider the fact that there is a spectrum of possible variations. In particular, there may be a variation that can make both horns of the dilemma work.

[28] A couple of clarifications about nudging. First, one might think, as some have suggested, that the evidence we have for the efficacy of nudging is shaky, at best (Bakdash & Marusich, 2022; Maier et al., 2022; Szaszi et al., 2022). Yet a study which survives these recent criticisms provides evidence that some nudges are efficacious (DellaVigna and Linos 2022). Second, it has been argued that such influences bypass at least some of our capacities, though maybe not exactly the same ones that bypassing theorists refer to. If, however, they do not, they might be more like the sorts of influences in the first version of *LWMI* (for discussion, see (Douglas, 2022; Levy, 2017, 2018, 2019; Schmidt 2019)).

[29] For brief discussions of how features of such cases can be modified to yield more or less substantial manipulation, see (Mele, 2019, pp. 37–8, 130–3).

Once we start to consider the whole spectrum of cases, they will get messy, as will our judgments about them. Yet, just as the significance of the influence from the signals can vary, so can its effects on Plum's responsibility. Suppose one thinks that Plum in *SMI* is not responsible for killing White, and that Plum in *WMI* is just as responsible as a typical agent in a deterministic universe would be. One might still think that, were the influence to operate via bypassing, and result in a somewhat more robust change than what we find in *WMI*, yet one that is not as substantial as in *SMI*, Plum could be responsible for killing White, yet still *less* responsible than a typical agent in a deterministic universe would be. For many of the cases on the spectrum, this might be precisely what is going on.

A bypassing view, I suggest, may well be capable of accounting for this. Such a view can go beyond a necessary condition on responsibility for an action and suggest ways in which, as a result of bypassing, an agent's responsibility for some action or other might be mitigated.[30] For instance, one might suggest, after pointing to at least some of the features mentioned above, that if an agent undergoes a change to her attitudes via bypassing, then to the extent this results in a more or less severe change, or a more or less severe mitigation of the relevant capacities and opportunities, her responsibility for some of her actions is mitigated.[31]

Thus, just as the influences in variations of *Machine Induction*, as well as cases of manipulation, can come in various forms, and involve changes of different severity, so might their negative effect on agents' responsibility; responsibility might be eliminated in some cases, it might be mitigated in others (to varying degrees), and in others still, have no effect. And bypassing views can plausibly account for this.

What if, again, we eliminate the assumption that this is the first time that the signals influence Plum? That is, what if we further stipulate that he has been undergoing a variety of signals, along this spectrum, throughout his lifetime? For the reasons mentioned above, this sort of case will be difficult to assess, both in terms of applying the views, and in making judgments about such a case. And once we begin to consider a whole spectrum of different influences, and we say that they occur throughout Plum's entire lifetime, this opens up the possibility of a vast array of combinations of influences, and thus a vast array of cases. Consequently, there is unlikely to be a "one-size-fits-all" response from the bypassing theorist here.

In relation to this, it is important to point out a further feature of bypassing views. Even in cases where the agent undergoes a radical reversal via bypassing, a change like that found in *SMI*, the agent may come to be responsible for actions that issue from the attitudes acquired via bypassing again. This can happen after having had the relevant sort of opportunity to either shed the attitudes, learn to resist them, or integrate them into their system of values.[32] This can also apply to cases in which

[30] One might, instead, think that it is blameworthiness and praiseworthiness that comes in degrees. For some discussion, see (Coates, 2019).

[31] Cyr and McKenna both suggest that, even if one takes a hard-line with respect to manipulated agents, one can still hold that they are *less* responsible, or responsible for *less* things, (Cyr 2020, pp. 2392–3; McKenna 2016, p. 93). Some of their points could, it seems, be employed by a bypassing theorist as well.

[32] For discussion, see (De Marco, 2021, pp. 12–5, 2022, pp. 1960–1; Fischer, 2012, pp. 203–4; Fischer & Ravizza, 1998, Chap. 11; McKenna, 2016; Mele, 2020, p. 3149).

the influences merely mitigate, but do not fully undermine, responsibility for actions issuing from the newly acquired or modified attitudes. Thus, although some of these influences may mitigate responsibility for some actions, responsibility for these actions need not be mitigated in perpetuity.

4.3 A lifetime of weak manipulation[33]

Finally, one might worry that it is not simply a variation of *Machine Induction* that threatens bypassing views. Rather, it is the comparison of (1) a version of *Machine Induction* on which bypassing theories would not get the result that Plum lacks responsibility for killing White, and (2) an analogous case that involves a manipulator, in which it seems that Plum *does* lack responsibility for killing White. This is the sort of concern that Liu may be raising with the second horn of the dilemma.

What might such cases look like? Recall *LWMI*. In this case, Plum is influenced by signals from the machine throughout his lifetime, yet the influences are of the sort that appear in *WMI*. The individual influences do not result in new attitudes that Plum is compelled to possess, they are not the result of a radical reversal, he retains control over his actions, if they are like the case of Carl, then they are easy to resist, etc. This is also true of the influence from the signal that Plum receives shortly before deciding to kill White, and of his action of killing White. A bypassing view according to which Plum in *WMI* and Carl are responsible for their relevant actions would seem to suggest that Plum in *LWMI* is also responsible for killing White. As I suggested above, Plum in *LWMI* might not be subject to influences that are significantly different from those we already encounter in everyday life. If one does not think that those influences undermine our responsibility for our behavior, then this result does not seem problematic.

However, compare this case to what we can call *A Lifetime of Weak Manipulation* (*LWM*). This case is like *LWMI* in that it involves the same sorts of influences on Plum throughout his life, yet this time, there is a manipulator behind these influences. After a long series of these minor interventions, Plum eventually receives the final weak influence, decides to kill White, and does so. Would claiming that Plum is responsible for killing White in *LWM* involve making a counterintuitive claim, and thus constitute a hard-line response? This will depend on what intuition one has about this case, and I am not sure what intuition people will have.

In thinking about this case, at this abstract level of description, it may be easy to simply focus on the fact that Plum is being slightly nudged, throughout his life, to end up killing White. Yet a focus on this can distract from the mildness of these influences, and the fact that for every one of these influences, it is false that it produces an attitude that Plum is compelled to possess and it is false that the change in attitudes is part of a radical reversal. And it can distract from the role that Plum's agency played throughout: when he acts on these influences, he has control over his behavior, and if they are like the case of Carl, the resulting attitudes are easy

[33] I am grateful to an anonymous reviewer for suggesting this sort of case.

to resist. Thus, one way to think of this case is as one in which there is a series of influences which eventually lead Plum to kill White. But this might distract from another, equally adequate description of the case, on which a series of choices by Plum, over which he had control, and none of which were the results of attitudes that were significantly modified by these signals, resulted in his killing White.[34] If we are further considering a case in which, as in the first versions of *WMI* and *LWMI*, these influences do not bypass, and thus engage with, Plum's capacities for control over his mental life, it seems difficult to think of Plum as lacking responsibility for killing White, while also thinking that ordinary agents *are* responsible for their actions.

Finally, recall a point made above about Plum in *LWMI*. There, I suggested that the influences from these signals may not be relevantly different from influences we already encounter in our everyday lives via, for instance, various marketing techniques, product design, or nudges. Yet many of these tend to be produced by agents attempting to influence our attitudes and behavior, and they often involve multiple influences geared towards the same change in attitudes—e.g., acquiring or increasing a desire for a certain product or candidate—or toward a particular behavior—e.g., purchasing that product or voting for that candidate. That is, we are already likely subject to a variety of fairly localized versions of *LWM* intended to result in a variety of behaviors.

Though I do not mean to suggest that these points ought to convince everyone that Plum *is* responsible for killing White in *LWM*, they can help to show why a judgment about this might not be a clear one, nor an easy one to come to.

5 Concluding thoughts

Determining which form of the soft-line reply is preferable is going to be a difficult matter, and involve an assessment of not just the cases considered here, but a variety of others.[35] What I have argued for here is not that the bypassing approach is preferable to a manipulator-focused approach, all things considered; the conclusions are much more limited, and mainly restricted to cases of *Machine Induction*. Where

[34] One might worry that, if a set of these signals occur quite rapidly, and have influences that add up to something more significant relatively quickly, then this might change things. Yet there are two points to keep in mind here. First, in such a case, this might be more like a case involving influences further on the spectrum. Second, if Plum does not have the relevant opportunities to shed, learn to resist, or incorporate these attitudes into his system of values, then *LWM* will be significantly different from *LWMI* in a way that bypassing views can track.

[35] Another important sort of case are cases of original design; e.g., cases in which the intervener intervenes before the victim is a full-blown agent, or even before he comes into existence (Mele, 2006, p. 188; Pereboom, 2014, p. 77). If such agents are not responsible for the relevant actions, bypassing views do not have the resources to account for this. And for some of the most recent manipulator-focused accounts—e.g., Usher's (Usher 2020) and Deery & Nahmias's (2017)—another sort of case is that of a lucky manipulator (Pereboom and McKenna, 2022, pp. 192–3; Tierney Forthcoming, n. 7; Tierney & Glick, 2018, pp. 958–9, n. 7). Such views do not seem capable of accounting for the manipulated agent's lack of responsibility in these cases. Further, depending on how one interprets his view, Usher's may not be able to account for our intuition that a subject of original design lacks responsibility for the relevant action (Tierney & Glick, 2018, p. 958, n. 6; Usher, 2020, p. 320).

does this leave us, then, with respect to the dialectical purpose and force of the case(s) of *Machine Induction*? The arguments in Sect. 3 suggest that Liu's dilemma does not present a problem for the bypassing theorist, and the case of *Machine Induction*, on either interpretation, does not give the bypassing theorist a reason to adopt a manipulator-focused view.

In *SMI*, Plum undergoes a radical reversal, such that killing White, something he was not previously capable of acquiring a desire for, is now an attractive option for him, etc. If one thinks that this Plum is not responsible for killing White, then one might have reason to adopt a bypassing view, insofar as such a view has the resources to account for Plum's lack of responsibility. Manipulator-focused views, on the other hand, cannot account for Plum's lack of responsibility, given the absence of a manipulator.[36] Further, if one similarly thinks that Plum is not responsible for killing White in *LSMI*, which involved a lifetime of such influences, then similar points apply; bypassing views can account for Plum's lack of responsibility for killing White, whereas manipulator-focused views cannot. Thus, not only can the bypassing theorist account for Plum's lack of responsibility for killing White without rejecting compatibilism—and thereby avoid the implication of Liu's first horn—this case of machine induction would seem to remain an issue for the manipulator-focused views.

In *WMI*, with or without bypassing, the influence from the machines is not significantly different than one we might expect from "a momentary alteration in attention due to bad digestion…or an unexpected remark about one's abusive father" (McKenna, 2017, p. 579). With respect to this case, the bypassing theorist can accept the no-difference thesis, and accept that Plum is responsible for killing White. They can then argue that accepting that Weakly Manipulated Plum—in a case just like *WMI*, yet which involves a manipulator—is responsible for killing White is either not too hard of a line to take, or not a hard line at all.

With respect to *LWMI*, which involves such influences throughout his entire life, I have suggested that the bypassing theorist might still hold that Plum is responsible for killing White. Given the lack of a manipulator, this would also seem to be the position of one who holds a manipulator-focused view. Thus again, we do not have a case on which the two differ. Finally, with respect to *LWM*, which is like *LWMI* yet the signals are sent by a manipulator, I have suggested that things get murkier. Assuming that there are no differences that are relevant to bypassing views between *LWM* and *LWMI*, it would seem that bypassing views would treat the two Plums in the same way.

If one thinks that Plum in *LWM* is responsible for killing White, then the bypassing view does not face an issue. If, instead, one thinks that Plum is not responsible for killing White in this case, then one might think that this case poses a problem.

[36] Again, the possible exception being Deery and Nahmias's view, who show how their view can account for lack of responsibility in at least some cases which lack a manipulator (Deery & Nahmias, 2017, pp. 1272–3). Liu pushes back on this, as it applies to a variation of *Machine Induction* (Liu, 2022, pp. 539–43), and Deery and Nahmias discuss a potentially similar case, suggesting that their view might not account for the agent's lack of responsibility for the relevant action (Deery and Nahmias, 2017, p. 1268, n. 12).

Finally, if one does not have a clear judgment about this case, then it may not be all that useful for deciding between the views.

Ultimately, insofar as we assess soft-line views partly in terms of whether they align with our judgments about cases, we will need to consider a variety of other cases as well. And as things stand, every view has some problematic cases they cannot deal with.[37] As I have argued here, although *LWM* may be such a case for bypassing views, depending on one's take on it, *SMI* and *LSMI* are not; and the latter would seem to remain problematic for manipulator-focused views.

Acknowledgements For discussion and comments on a previous draft of this paper, I would like to thank Alfred Mele, Taylor Cyr, and Thomas Douglas. I would also like to thank an anonymous reviewer for this journal who gave extensive and thoughtful comments which helped to substantially improve this paper.

Funding I would like to thank the European Research Council [ERC Consolidator Award 819757 (ProtMind)] for their financial support.

Declarations

Conflict of interest The author declares no conflict of interest.

References

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.

Arpaly, N. (2006). *Merit, meaning, and human bondage: An essay on Free Will*. Princeton University Press.

Bakdash, J.Z., & Marusich, L.R. (2022). Left-truncated effects and overestimated meta-analytic means. *Proceedings of the National Academy of Sciences*, *119*(31), e2203616119. https://doi.org/10.1073/pnas.2203616119

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

Barnes, E. C. (2015). Freedom, creativity, and manipulation. *Noûs, 49*(3), 560–588.

Coates, D. J. (2019). Being more (or less) blameworthy. *American Philosophical Quarterly, 56*(3), 233–246.

Cyr, T. W. (2020). Manipulation and constitutive luck. *Philosophical Studies*, *177*(8), 2381–2394.

De Marco, G. (2021). Historical moral responsibility and manipulation via deletion. *Erkenntnis*. Accessed 26 May 2021

De Marco, G. (2022). Nonconsensual neurocorrectives, bypassing, and free action. *Philosophical Studies*, *179*(6), 1953–1972.

Deery, O., & Nahmias, E. (2017). Defeating manipulation arguments: interventionist causation and compatibilist sourcehood. *Philosophical Studies*, *174*(5), 1255–1276.

[37] See n. 35 above.

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica, 90*(1), 81–116. https://doi.org/10.3982/ECTA18709
Demetriou, K. (2010). The soft-line solution to Pereboom's four-case argument. *Australasian Journal of Philosophy*, *88*(4), 595–617.
Dennett, D. (1984). *Elbow room: The varieties of free will worth wanting*. MIT Press.
Douglas, T. (2022). If Nudges treat their targets as rational agents, nonconsensual neurointerventions can too. *Ethical Theory and Moral Practice*, *25*(2), 369–384.
Fischer, J. M. (2006). *My way: Essays on moral responsibility*. Oxford University Press.
Fischer, J. M. (2012). *Deep control: Essays on free will and value*. Oxford University Press.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.
Frankfurt, H. G. (2002). Reply to John Martin Fischer. *The contours of agency: Essays on themes from Harry Frankfurt* (pp. 27–31). MIT Press.
Haji, I. (1998). Moral appraisability: Puzzles, proposals, and perplexities. Oxford University Press.
Haji, I. (2010). The inauthentic evaluative schemes of psychopaths and culpability. *Responsibility and psychopathy: Interfacing law, psychiatry and philosophy* (pp. 261–281). Oxford University Press.
Haji, I., & Cuypers, S. (2001). Libertarian free will and CNC manipulation. *Dialectica*, *55*(3), 221–239.
Haji, I., & Cuypers, S. (2008). *Moral responsibility, authenticity, and Education*. Routledge.
Herdova, M. (2021). The importance of being Ernie. *Thought: A Journal of Philosophy, 10*, 257–263.
Kaiserman, A. (2021). Reasons-sensitivity and degrees of free will. *Philosophy and Phenomenological Research*, *103*(3), 687–709.
Kearns, S. (2012). Aborting the zygote argument. *Philosophical Studies*, *160*(3), 379–389.
King, M. (2013). The problem with manipulation. *Ethics, 124*, 65–83.
Levy, N. (2017). Nudges in a post-truth world. *Journal of Medical Ethics*, *43*(8), 495–500.
Levy, N. (2018). Nudges to reason: Not guilty. *Journal of Medical Ethics, 44*(10), 723–723. https://doi.org/10.1136/medethics-2017-104639
Levy, N. (2019). Nudge, nudge, wink, wink: Nudging is giving reasons. *Ergo, an Open Access Journal of Philosophy*, *6*.
Liu, X. (2022). Manipulation and machine induction. *Mind*, *131*(522), 535–548.
Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., & Wagenmakers, E. J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), e2200300119. https://doi.org/10.1073/pnas.2200300119
Matheson, B. (2016). In defence of the four-case argument. *Philosophical Studies*, *173*(7), 1963–1982.
McKenna, M. (2008). A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research, 77*(1), 142–159.
McKenna, M. (2016). A modest historical theory of moral responsibility. *The Journal of Ethics, 20*(1), 83–105.
McKenna, M. (2017). Manipulation arguments, basic desert, and moral responsibility: Assessing Derk Pereboom's free will, agency, and meaning in life. *Criminal Law and Philosophy, 11*(3), 575–589.
Mele, A. (1995). *Autonomous agents: From self-control to autonomy*. Oxford University Press.
Mele, A. (2005). A critique of Pereboom's "Four-Case Argument" for Incompatibilism. *Analysis*, *65*(1), 75–80.
Mele, A. (2006). *Free will and luck*. Oxford University Press.
Mele, A. (2008). Manipulation, compatibilism, and moral responsibility. *The Journal of Ethics, 12*(3), 263–286.
Mele, A. (2019). *Manipulated agents: A window to moral responsibility*. Oxford University Press.
Mele, A. (2020). Moral responsibility and manipulation: On a novel argument against historicism. *Philosophical Studies, 177*(10), 3143–3154.
Mickelson, K. (2019). The manipulation argument. In K. Timpe, M. Griffith, & N. Levy (Eds.), *The Routledge companion to free will* (pp. 166–178).
Pereboom, D. (2014). *Free will, agency, and meaning in life*. OUP Oxford.
Pereboom, D. (2017) A defense of free will skepticism: Replies to commentaries by Victor Tadros, Saul Smilansky, Michael McKenna, and, & Alfred, R. Mele on free will, agency, and meaning in life. *Criminal Law and Philosophy*, *11*(3), 617–636.
Pereboom, D., & McKenna, M. (2022). Manipulation arguments against compatibilism. *The Oxford handbook of moral responsibility* (pp. 179–200). OUP.
Schmidt, A. T. (2019). Getting real on rationality—behavioral science, nudging, and public policy. *Ethics, 129*(4), 511–543.

Sekatskaya, M. (2018). Double defence against multiple case manipulation arguments. *Philosophia*. Accessed 11 June 2019

Shaw, E. (2014). Direct brain interventions and responsibility enhancement. *Criminal Law and Philosophy*, *8*(1), 1–20.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., et al. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, *119*(31), e2200732119. https://doi.org/10.1073/pnas.2200732119

Takasaki, S. (2021). An argument against the methodology of the Manipulation Argument. *Review of Analytic Philosophy, 1*(1), 51–61.

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The final edition*. Allen Lane.

Tierney, H. (Forthcoming) (Ed.). The future of the causal quest. In *Blackwell* companion to free will. Wiley.

Tierney, H., & Glick, D. (2018). Desperately seeking sourcehood. *Philosophical Studies*. https://doi.org/10.1007/s11098-018-1215-3

Usher, M. (2020). Agency, teleological control and Robust Causation. *Philosophy and Phenomenological Research, 100*(2), 302–324.

Vihvelin, K. (2013). *Causes, laws, and free will: Why determinism doesn't matter*. OUP.

Waller, R. R. (2014). The threat of effective intentions to moral responsibility in the Zygote argument. *Philosophia, 42*(1), 209–222.

Yaffe, G. (2003). Indoctrination, coercion and freedom of will. *Philosophy and Phenomenological Research, 67*(2), 335–356.