Philos Stud
https://doi.org/10.1007/s11098-019-01292-2

The agential perspective: a hard-line reply to the four-case manipulation argument

Sofia Jeppsson[1]

Abstract One of the most influential arguments against compatibilism is Derk Pereboom's four-case manipulation argument. Professor Plum, the main character of the thought experiment, is manipulated into doing what he does; he therefore supposedly lacks moral responsibility for his action. Since he is arguably analogous to an ordinary agent under determinism, Pereboom concludes that ordinary determined agents lack moral responsibility as well. I offer a hard-line reply to this argument, that is, a reply which denies that this kind of manipulation is responsibility undermining. I point out that fully fleshed-out manipulated characters in fiction can seem morally responsible for what they do. This is plausibly because we identify with such characters, and therefore focus on their options and the reasons for which they act rather than the manipulation. I further argue that we ought to focus this way when interacting with other agents. We have no reason to trust the incompatibilist intuitions that arise when we regard manipulated agents from a much more detached perspective.

Keywords The four-case manipulation argument · Compatibilism · Incompatibilism · Moral responsibility · Intuitions

1 The four-case manipulation argument

Derk Pereboom's *four-case manipulation argument* is one of the most influential arguments for incompatibilism in the moral responsibility debate. The argument's four cases feature Professor Plum, who murders Ms White for selfish reasons

✉ Sofia Jeppsson
sofia.jeppsson@umu.se

[1] Department of Historical, Philosophical and Religious Studies, Umeå University, 901 87 Umeå, Sweden

Published online: 03 April 2019

(Pereboom 2001: Chap. 4, 2014: Chap. 4). Plum is stipulated to fulfil every popular compatibilist moral responsibility condition; he is reasons-responsive (Fischer and Ravizza 1998), his first-order will conforms to his second-order volitions (Frankfurt 1971), he is capable of taking moral as well as prudential reasons into account (Wallace 1994), he acts in character (Hume 1739/1978) and so on. Plum is, however, manipulated into acting the way he does. In case 1, a team of neuro-scientists manipulate him from moment to moment in order to cause him to murder White. This is supposed to elicit a strong non-responsibility intuition in us readers, and prompt us to judge that Plum lacks moral responsibility in this situation. In case 2, Plum was programmed at birth by a team of neuro-scientists to become a fairly egoistic person, determined, when finding himself in circumstances where he stands to gain a lot from doing so, to kill White. Pereboom writes that it can hardly be morally relevant whether the scientists manipulate Plum from a temporal distance or not, so if he lacks moral responsibility in the first case, he does so in the second one as well. In case 3, the scientists and their science-fiction technology are replaced by a community who brings Plum up to be rationally egoistic, since they value this trait. Plum ends up egoistic enough to be determined to murder White when circumstances are such that he stands to gain a lot from the murder. Once again, Pereboom writes, it cannot be morally relevant whether Plum is manipulated through science-fiction-like or more mundane means. If Plum lacked moral responsibility in the previous cases, he does so in case 3. Finally, we arrive at case 4, ordinary determinism. It cannot matter, says Pereboom, whether the manipulation of Plum is made by other agents or if he is determined to do what he does by nature. Thus, Plum lacks moral responsibility in case 4 as well.

Now, there has been some debate about case 1. In Pereboom's first version of this case, Plum was said to be manipulated by neuro-scientists from moment to moment by "radio-like technology" (Pereboom 2001: 112–113). It has been argued that in this scenario, Plum is not integrated enough to be a real agent (Demetriou 2010; Fischer 2004: 156; Mele 2005: 78; Baker 2006: 320). In response, Pereboom produced a new and significantly weaker version of case 1, where the scientists merely push a button making Plum momentarily extra selfish. Pereboom argues that this kind of manipulation does not destroy agency, since we are often caused to be momentarily extra selfish in real life by events such as stubbing a toe on a table leg or hearing that our favourite team just lost, and no one thinks this robs us of agency (Pereboom 2014: 76–77). The problem with the new case 1 is that while agency preserving, it might fail to evoke the strong non-responsibility intuition that the argument relies on. Matheson (2016) argues that the first version of case 1 was actually superior, and the critique of it flawed. I will not try to resolve this debate here, although I will return to case 1 later in the paper. For the moment, I will be content to note that Pereboom's argument remains quite strong even if we start with case 2, where Plum was programmed at birth to become the person he is.

Michael McKenna gives what he calls a "hard-line reply" to *the four-case manipulation argument* (2005, 2008, 2013). Counter arguments according to which there *is* a morally relevant difference between ordinary determined agents and manipulated ones he labels "soft-line replies". McKenna does not really believe in this strategy; for every supposedly morally relevant difference between

manipulation and ordinary determinism pointed out by compatibilists, Pereboom and other incompatibilists can simply come up with new and improved manipulation cases that eliminate these differences. A hard-line reply, on the other hand, denies that this kind of manipulation undermines moral responsibility. McKenna suggests that we challenge our non-responsibility intuitions about the early cases by beginning with case 4, ordinary determinism, and focus hard on Plum's agential qualities. This will elicit compatibilist intuitions and prompt us to judge him morally responsible for the murder, or at least to remain agnostic about the matter. The compatibilism or agnosticism can then be carried over to case 3, 2 and 1.[1]

I am sympathetic to McKenna's hard-line reply, but it does not quite work for me. If I start with case 4, my intuitions about *that* case are at least not obviously incompatibilist. But when moving through case 3 and arriving at case 2, I hit a kind of intuitive wall. On reading Pereboom's description of case 2, I strongly feel that Plum cannot be morally responsible for what he does; after all, he merely follows his program (see also Pereboom 2005: 242). I cannot overcome the non-responsibility intuition by simply repeating what Pereboom writes about Plum's agential qualities and try to focus on those. Interestingly, however, this is not *always* the case when I partake of stories featuring characters who have been programmed to do what they do—and I strongly suspect that in all this, I am not alone.

2 Shifting intuitions

Science fiction features a whole bunch of robots and artificial intelligences that we intuitively consider morally responsible for what they do. One particularly interesting case concerns the Hologram Doctor from the TV show Star Trek Voyager. Just like Plum in case 2 (hereafter, Plum 2), the Doctor does what he does because his program interacts with environmental influences. The only real difference between the two characters is the amount of information and foresight that their creators possessed. The neuro-scientists who created Plum 2, we are told, *intended* for him to kill White; they must thus possess a next to god-like omniscience regarding all the experiences that Plum 2 will gather through his life and the situations he will find himself in, in order to make him such that he, while reasons-responsive, acting in character and so on, will be determined to kill White when standing to gain a lot from doing so.[2] Lewis Zimmerman, the engineer who creates the Hologram Doctor in Star Trek, does not know any more about the future than any other person, and has thus neither foreseen nor intended for the Doctor to perform all the particular actions he ends up doing. Still, Plum 2 and the Doctor resemble each other in that both do what they do as a result of their program interacting with their experiences and environmental influences in a generally rational and reasons-responsive way.

[1] Or, rather, it can be carried from case 6 to case 1, since McKenna suggests two extra intermediate steps, making ordinary determinism case 6 rather than case 4.

[2] At least this is the case in the scenario described in Pereboom (2014: 77); the mention of an intention is not present in every version of the manipulation argument.

It often remains vague to what extent a robot or artificial intelligence in a science fiction story is programmed to do what he does; the Hologram Doctor is an interesting case since he *explicitly* states (episode 5, season 1) that he follows a program. He worries, for that reason, that he might not be praiseworthy for having saved a man's life. Kes, a humanoid alien, tries to cheer him up. She first says that it was still *him* who saved the man's life, and proceeds to point out the ways in which the Hologram Doctor is similar to a regular, humanoid MD: the Doctor as well as regular MD:s gradually evolve and learn through experience. This little exchange between the two characters is reminiscent of the dialectic between Pereboom and McKenna; Pereboom sees manipulation or programming as responsibility undermining, whereas McKenna urges us to focus on the ways in which Plum resembles a regular person. However, although McKenna's hard-line reply fails to really move me, what the character Kes says in the TV show makes perfect sense to me; I intuitively feel that the Hologram Doctor is praiseworthy for saving the man's life, and overall morally responsible for what he does, despite following a program. But why is that?

Here is my hypothesis: Pereboom's descriptions of Professor Plum draws attention to the way in which he resembles a highly sophisticated vending machine. Insert a coin, out comes a candy bar—insert the scientists' program and the right environmental factors, out comes a murder. We are told that Plum is reasons-responsive, has a first-order will which corresponds to his second-order volitions and so on—but all those agential features seem like nothing but cogs in a sophisticated machine. Even though we are *told* that Plum is a fully-fledged rational agent, it is hard to fully *appreciate* this fact when he is described the way he is. (This is no critique of how Pereboom frames the case; thought experiments are necessarily brief in their descriptions. You cannot insert an entire novel into a philosophy paper or book.) A fictional character like the Hologram Doctor, who is presented to us as a fully fleshed-out agent, is perceived differently. As viewers, we are invited to identify with him and the other characters on the show, and to see matters from their point of view. When we do this, our focus naturally moves away from the fact that the Doctor was programmed to do what he does. After all, most of the time, the Doctor faces various decisions and has to decide what to do by weighing his reasons just like the rest of us. When we are invited to identify with him and see things from his point of view, we, too, focus on his reasons, rather than the fact that an engineer once provided him with a program that ultimately determines what he does. When we focus on his reasons for action and which one he ought to choose, we do not get the non-responsibility intuitions that a briefly described thought experiment gives us.

The fact that this shift in focus happens pretty naturally when we identify with a programmed character in a piece of fiction does not entail, of course, that we should trust the resulting intuitions. Pereoom (2005) writes, in response to McKenna, that we ought to focus on both Plum's agency and his manipulation, and that the way he describes the thought experiment draws attention to both. The compatibilist cannot say that we ought to focus exclusively on Plum's agency, since that would be question-begging. Pereboom could thus agree that the Hologram Doctor seems morally responsible to us when we watch Star Trek Voyager, but dismiss this

intuition as a misleading one, caused by us forgetting that the Doctor was programmed to do what he does when we, e.g., watch him trying to save someone's life.

I will argue that making responsibility judgments about a character manipulated in this agency-preserving manner while ignoring the manipulation is not only natural but justified. When holding others responsible, we ought to place ourselves in their shoes and regard them from the same point of view as they regard themselves. I will argue, in the sections that follow, that manipulated agents have reason to disregard their own manipulation when deciding what to do. Since they have reason to disregard it, so do we.

3 Plum 2's decision

Let us first take a look at Plum 2, and what things seem like from his perspective. Imagine that he finds himself in the following situation: He is the only heir of his elderly relative Ms White. He realizes that he has an opportunity to sneak into her house and smother her with a pillow, sneak out again, and have the whole thing looking like death from natural causes. If need be, some friends of his are even ready to provide him with an alibi for the time of Ms White's death. He is thus pretty certain that he will get away with the murder and earn a lot of money. If he does nothing, he will eventually inherit the money anyway, but for various reasons he desperately needs money *now*. Still, Plum does not take murder lightly (he is not, after all, a psychopath incapable of appreciating moral reasons for action). Can he really *kill* for money? Can he really rob White of the years she still, presumably, has ahead of her?

If Plum 2 does not *know* that a team of neuro-scientists created and programmed him, he can obviously not take this fact into account when deciding what to do. Pereboom presumably intended for Plum 2 to be ignorant of his manipulated condition, just like people in general, he claims as he is quoting Spinoza, are ignorant of our lack of control over what we do, and therefore mistakenly think ourselves free and morally responsible (Pereoom 2005: 242). Suppose, however, that Plum *does* know that he is manipulated. What, exactly, is he supposed to *do* with this information? Plum is selfish, but not completely unmoved by morality. The pro tanto reasons (I here use "reason" in a motivational sense; I do not take a stand on whether Plum could have objective reasons for murder) that Plum has to weigh in order to decide what to do is thus on the one hand that he stands to gain a lot from White's death, on the other hand that murder is immoral. He thus has some hard deliberation to do regarding morality versus money; but pondering his manipulated state simply cannot help him decide one way or the other.

We can imagine Plum thinking that whatever he does, his decision will be a sham, so he might just as well do nothing. But if he thinks this he will, in effect, have chosen *not* to kill Ms White (and not getting the money). Furthermore, in that case, he can still be certain that this is what the program and the environment caused him to do, and so his supposed non-decision ends up being as much of a sham as any other decision he could have made. Plum simply cannot make his choice any less of

a sham by deciding to do nothing. Sham or no sham, he has to decide what to do. And in order to decide, he has to weigh his reasons pro and con each option.

Plum might of course think to himself that he does not *really* have two options to choose between. He knows that he follows a program, and that given his program and the environment in which he finds himself only one action is possible. But pondering this fact is just as useless as thinking of his upcoming decision as a sham; doing so will not bring him any closer to a decision. Furthermore, it is true that Plum has more than one option *in a sense*; whether he will murder White or not depends on his deliberation, choice and intention. His program is not agency bypassing. In order to decide what to do, Plum must weigh his reasons pro and con each (in the *above-mentioned sense* open) option.

Of course, given peculiar enough preferences, values and beliefs, almost anything can be a reason for any action. We might imagine that Plum is worried about the possibility of performing a seriously wrongful action for which he will be responsible (perhaps, for instance, Plum believes that if he does so, he will go to Hell when he dies), but is far less worried about the possibility of blameless wrongdoing. Plum furthermore *believes* (whether he is right or not can be set aside for the moment) that if he murders White while manipulated, he is not morally responsible for doing so, but if he murders her unmanipulated, he is. The fact that he *is* manipulated thus supplies him with a reason to commit the murder and take the money. In this scenario, Plum thus takes the fact of his manipulation into account when deciding what to do—in a sense. In a different sense, he still has to ignore it. He must still think of "murder White" and "not murder White" as two options between which he has to choose—he cannot get hung up on the fact that if he ends up thinking "if I murder White, I get all the money, and I won't go to Hell, so I should murder her" he was manipulated into thinking *this,* whereas if he ends up thinking "murder is, after all, a horrible and morally wrong thing to do, so I shouldn't murder her" he was manipulated into thinking *that.* Even when he, in this sense, takes manipulation into account when deciding what to do, he must think of murdering White and abstaining from murder as *two options* between which he has to *choose.*

It is thus not merely the case that Plum mistakenly thinks that he has a choice because he is unaware of being manipulated, just like ordinary people (according to Pereboom) mistakenly thinks that we have real choices about what to do because we are unaware of all the factors that causes us to act. Even in my tweaked scenario, where Plum knows about the manipulation, the manipulation is largely *irrelevant* from his point of view. He has *reason* to disregard it, and to instead think of "murder White" and "not murder White" as two possibilities, between which he has to choose by weighing his pro tanto reasons pro and con each option.

Someone might object that Plum disregards his manipulation in this scenario because although he knows that he is manipulated, he still retains a little bit of Spinozic ignorance; he does not know exactly what he is manipulated to do. But what if the scientists told him, not just that they have manipulated him into doing whatever he ends up doing, but that they have manipulated him into *murdering White?* Well, what happens in that scenario is a complicated and controversial question; however, it is ultimately not relevant for the present discussion.

Whether it is even in principle possible to accurately and reliably predict someone's behaviour while also continuously informing him about the predictions made is controversial (see Bok 1998: 81–87 for the argument that it is not). If it were possible, it is far from obvious how this kind of knowledge would affect agents; perhaps not much at all, because as long as our actions depend on our thoughts and intentions we will still have to think things through and decide what to do (Korsgaard 1996b: 94–96; Jeppsson 2016: 1235). Perhaps there is no reason to care about one's manipulation, even when one knows where it will lead.

But let us assume, for the sake of argument, that if the scientists told Plum everything he was manipulated into doing before he did it, his lack of "epistemic openness" would stop him from deliberating about what to do (Kapitan 1986; Pereboom 2008). Unable to deliberate, he cannot make choices or decisions either; he ends up merely going through the motions of actions he knew he was destined to perform anyway. Plum in this scenario hardly seems like a real agent any longer, and is, for that very reason, arguably not morally responsible for murdering White. However, even if this version of Plum lacks moral responsibility due to lacking agency, we cannot conclude that regular determined people who *have* agency therefore lacks moral responsibility as well. We must distinguish between a scenario where additional knowledge reveals to an agent that he lacked moral responsibility all along, and a scenario where too much knowledge *destroys* his previously held moral responsibility due to destroying his agency. For all that I have said about the perfectly knowledgeable version of Plum in this paragraph, the latter rather than the former might be true of him.

Let us therefore lay the *fully* informed version of Plum aside, and move back to the version of Plum 2 who merely knows that he is manipulated, but not what he will do. In order to decide what to do, he must focus on his reasons for action, and set aside the fact that whatever he ends up doing will be the result of his manipulation. Even if he occasionally, like the Hologram Doctor, ponders the fact of his manipulation and asks himself whether he really can be morally responsible for what he does, he must set these thoughts aside in situations where a decision is to be made. And *we* tend to largely ignore manipulation as well when we follow such characters in fiction, identifying with them, seeing the choices they face from their point of view.

I will now argue, more controversially, that the same is true for Plum 1 as well, before discussing what all this implies for Plum's moral responsibility.

4 Plum 1's decision

Once again, Plum naturally cannot take the fact of his manipulation into account when considering what to do if he does not know that he is being manipulated. Let us therefore imagine that he knows about it.

When discussing moment-to-moment manipulation, it is crucial to remember that Pereboom has stipulated that murdering White is in character for Plum. If the scientists, on the contrary, radically change Plum when manipulating him, and Plum knows this in advance, the manipulation presents him with some peculiar problems.

Imagine, for instance, that Plum normally is a very nice and altruistic person. He now finds out that for a period of time (a day perhaps, or a week) scientists will directly manipulate his brain, causing him to be selfish, and to believe that grossly immoral acts can be justified if he stands to gain enough from them. This is clearly very important news for Plum. He cannot, in this situation, simply resolve not to act on any selfish beliefs and desires he will find himself with during the following day or week, since such a resolution will not be effective once the manipulation starts.

Depending on how we fill in the details more precisely, Plum might find himself in a position similar to that of a weak-willed person. A weak-willed person might know that it is not sufficient for her to, e.g., resolve to quit smoking in order to really do so. She knows in advance that her resolve will weaken when the urge for a cigarette grows strong enough. She might therefore set out in advance to make smoking difficult: She rids her apartment of cigarettes, tells everyone around her not to give her cigarettes if she asks for them and so on. Plum might similarly try in advance to make murdering White difficult after the manipulation sets in. If the change is extreme enough, however, Plum might have to resort to treating his manipulated future self as another agent altogether: If he wants to prevent his future manipulated self from doing bad things, he must go about this business in roughly the same way as he would with another agent with whom he could not directly communicate. He might, for instance, try to warn as many people as possible to stay away from him during his time as manipulated. He might even take extreme measures such as having another person lock him up in a jail cell and leave him there for as long as manipulation lasts. Still, it is not exactly the *manipulation* which Plum 1 needs to take into account when making up plans in this scenario, but the *change*—if he were to find out in advance that he would suffer some kind of strange but spontaneously appearing brain affliction for a week which causes people to be extremely selfish, he would similarly have to take this into account when making plans.

If we imagine, instead, that Plum 1 is either directly and consistently manipulated throughout his entire life, *or* manipulated during a briefer period of time but by scientists who constrain themselves to follow up on his plans, adhere to his normal values and preferences and so on (see McKenna 2008: 149–150), Plum does not really have reason to take the manipulation into account when deciding what to do. Barring a preference for murdering White *while manipulated* (which I have already covered in my discussion of Plum 2), Plum 1 must decide, just like a non-manipulated counterpart would, whether morality or money is more important. Plum 1 is in the same situation as Plum 2. Musings on the fact that whatever they end up doing, they will have been manipulated to do, cannot help them make up their minds; they have to consider their options and focus on their reasons for action in order to make up their minds and get anything done (see also Korsgaard 1996a: 162–163).

Plum 1 characters are less common in science fiction than Plum 2 ones. Still, barring radical enough changes, Plum 1 characters will have to choose what to do the same way as Plum 2 characters, or, for that matter, regular people do. Quite possibly, if we were to consume fiction where we were encouraged to identify with a Plum 1 character, we, too, would focus on his options and reasons for action and

feel the intuition that the character is morally responsible for what he does, at least as long as no drastic enough changes are involved.

When extreme enough changes *are* involved, Plum himself must, as already pointed out, regard his manipulated and unmanipulated self as different agents. Even though they might be the same person on standard theories of personal identity, there are clear breaks in agency, since unmanipulated Plum cannot make plans that manipulated Plum will carry through and the other way around (at least not regarding the murder; possibly, unmanipulated Plum could still plan to buy a loaf of bread the next day and have manipulated Plum carry out this plan). Manuel Vargas (2013: 279) and Ben Matheson (2014, forthcoming) have made the plausible suggestion that although this kind of manipulation does not rule out moral responsibility, it prevents moral responsibility from transferring from one state to the other. If Plum is normally altruistic, but manipulated during a limited period of time to be highly selfish (although still reasons-responsive and so on), we could reasonably judge that altruistic Plum is morally responsible for what altruistic Plum does, and selfish Plum is morally responsible for what selfish Plum does, but altruistic Plum is not morally responsible for the actions of selfish Plum, nor the other way around. My guess is that when partaking of a sufficiently fleshed-out story featuring this kind of case, Vargas and Matheson's suggestion would at least not strike us as particularly counter-intuitive.

Nevertheless, it is one thing to say that Plum himself, when considering whether to murder White, will disregard the fact that whatever he ends up doing, he was manipulated into doing that. We cannot simply conclude from this fact that we also ought to disregard this fact when making our moral responsibility judgments. More argument is needed.

5 How ought we to focus?

Pereboom argues that we ought to focus on *both* Plum's agency and the manipulation, and points out that he describes both these features in the thought experiment. Prima facie, this sounds perfectly reasonable; surely, our intuitions will be most trustworthy if we try to take *everything* into account. Depending on what we mean by "focusing on agency", however, I doubt that this is possible.

We can simultaneously keep in mind that Plum is manipulated and that the manipulation is of a sophisticated kind, causing him to respond to reasons for action and so on, rather than acting like a simple robot or machine. However, in doing so, we still keep our eyes on the causal chain leading up to his action, from a kind of detached third-person perspective. It is a very different matter to place oneself in Plum's shoes and consider what the situation looks like to *him,* as I have done (somewhat inadequately, as we need fully fleshed-out fiction in order to really see a situation from someone else's point of view) in the previous sections, and as we frequently do when watching TV shows, reading novels and so on. I will distinguish between regarding Plum from a *causal* and an *agential* perspective respectively. When taking up a causal perspective, we focus on what caused the agent to do what she does. When taking up an agential perspective, we focus on what the

deliberation, decision and action was like *to her*. This distinction is somewhat inspired by the traditional Kantian distinction between regarding people from a theoretical or practical perspective, but in order to avoid bringing on all the baggage from centuries of Kantianism and deal with all the radically different interpretations Kant's texts have given rise to, I prefer to coin new terms. (There are also comparisons to be made with Strawson's participant/objective distinction—I will come back to that later).

It is important to note that the kind of switch in perspective that I here describe does not entail switching *beliefs*. When I take up an agential perspective on Plum, and see his choice to murder White from his point of view, it is not the case that I *cease to believe* that he is manipulated into doing what he does; that there is a deterministic causal chain running from the scientists' programming him through various environmental influences that interact with the program and culminates in the murder of White. I still *believe* this; it is just not what I *focus* on. Conversely, when I take up a causal perspective on Plum, and focus on the scientists and their manipulation, it is not the case that I cease to believe that there is something that it is like to be Plum, or that Plum must focus on his reasons for action in order to decide what to do. I still believe this, but I do not focus on it. Even keeping one's beliefs intact, however, a shift in focus can affect one's intuitions. And we can legitimately ask how we *ought* to focus or which perspective we *ought* to take up, the agential or the causal one, in order to arrive at trustworthy moral responsibility intuitions.

Pereboom writes that it would be question-begging of the compatibilist to suggest that we simply focus on Plum's agency and ignore the manipulation. If we imagine someone taking Pereboom's own description and then using a black marker to blot out the parts of the text describing the manipulation, afterwards repeating only the "agency" parts of the description over and over again, finally announcing that she feels responsibility intuitions coming on—I agree, this would be arbitrary and question-begging. However, considering what the situation looks like *to Plum* is not the same thing as merely focusing on *fewer* features of the situation, but on *different* features. To say that we should focus on the causal chain and disregard what things are like for Plum is no less question-begging than the suggestion that we do the opposite.

We might try to take *everything* into account by simultaneously focusing on what Plum's decision is like for him and on the fact that the manipulation deterministically causes him to do what he does, or by at least flipping back and forth between perspectives. In reality, I think this will be hard to do; either the causal or agential perspective will tend to dominate. We will tend to either focus on how the choice seemed from Plum's point of view, the options he had to choose between and his reasons for action, or on the causal chain (including the manipulation) leading up to the murder, but not on both at the same time.

Furthermore, it is not even obvious that our intuitions become more reliable the more information we try to take in. It is possible that adding irrelevant information will disturb our intuitions rather than helping them. It has been shown, for instance, that when asking people whether a fat man should be pushed down from a bridge in front of a runaway trolley in order to save others, suggesting that he is either black or white will influence people's judgments, even though the participants of the

survey hold the explicit belief that race does not matter (Uhlmann et al. 2009). Additional, irrelevant information about the man's race distort people's moral intuitions. I am not saying that the case of Plum and White is comparable to this version of the trolley problem, just that we have no reason to believe that moral intuitions generally improve the more information we try to take into account. Our intuitions might very well be at their best when we focus on all the *relevant* information and nothing else—but if so, we return to the question of what is relevant; Plum's own agential perspective or the causal one.

My suggestion is that different perspectives and different ways of seeing people's choices are appropriate in different contexts. Suppose, for instance, that I am a neuro-scientist myself, and I study what happens in people's brains when they deliberate and make decisions. In this context, I ought to focus on the causal, neurological chains leading up to the decision, and if I happen to have Plum as my subject, on the manipulation, since this is my job. In most contexts, however, it is appropriate to try to see things from other people's point of view. To which *extent* we ought to do this varies, of course, depending on how close our relationship is. With close friends, friendship might obligate me to really put myself in their shoes and try to appreciate as fully as possible what things were like for them when they acted. But even everyday interaction with people whom I do not have a close friendship with requires taking up their point of view to *some* extent. Simple, everyday conversations often involve references to options and the reasons we have for them. My colleague at work might say, for instance "Should I go to the Indian place or the Thai place for lunch? What do you think?" Even answering such a simple question becomes difficult if I start to ponder that wherever he ends up, it was determined by the past and the laws of nature (if the world is deterministic) or he was programmed to do it (if my colleague is Professor Plum). In order to answer, I need not empathize deeply with him and feel what he feels and so on. But I must think of "going to the Thai place" and "going to the Indian place" as two options, and think of some pro tanto reason for at least one of them (maybe the Thai place is cheaper)—just like my colleague thinks of his options and his reasons.

Deep friendships as well as simple everyday interactions thus require that we often see things from the other agent's point of view rather than focusing on what caused him to act. Furthermore, if we were to consistently take up a causal perspective on other people, this would come off as cold, even demeaning. I *morally* ought not to regard myself as an agent while regarding other people as sophisticated vending machines. If I focus, in my interactions with Plum, on his manipulation, I might try to influence him in various ways and figure out which figurative buttons to push in order to get desired results. But seeing him this way precludes treating him with the respect we owe other people—at least if this perspective dominates. Just like the Hologram Doctor occasionally pondered his programmed state, we might occasionally ponder determinism—or manipulation, if we came to interact with people whom we knew were manipulated. But we normally have to set this aside as soon as we consider what to do, or when we are engaged in anything resembling normal interaction with other people. And when we set determinism/manipulation aside, people also seem morally responsible for what they do.

It should be noted that I do not claim that we have to take up a *participant stance* towards other people in the full Strawsonian sense. Strawson (1962/2013) famously argued that we can take up either a participant or an objective stance. The former is, Strawson claimed, absolutely necessary for ordinary, adult human relationships. When we take up this stance towards others, we care about whether their actions express an ill, indifferent or good will towards us, and when the former, we react with resentment, indignation or anger. We hold people responsible and blame them for wrongdoing because this is so deeply ingrained in us that we cannot give it up, and even if we could, doing so would be undesirable, seeing as reacting against wrongdoing in this way is essential to normal relationships. Paul Russell (2010) argues that even if we came to interact with manipulated agents like Plum in the low number cases, we ought to take up a participant stance towards them (although the people responsible for the manipulation ought not to do so). However, whether the reactive attitudes of resentment, indignation and anger really are necessary for normal, adult human relationships is highly contested; perhaps we could replace these attitudes with calmer and friendlier ones of, e.g., sadness and disappointment, and perhaps our relationships would be the better for it (see, for instance, Pereboom 2014; Sommers 2007; Milam 2016).

My claim is much weaker than Strawson's. Strawson claimed that normal interactions with other people requires us to react with resentment and related attitudes to wrongdoing, but I take no stand on this issue in this paper. I merely claim that we cannot have normal interactions with other people unless we mostly see their choices from their point of view, rather than focusing on the causal chains behind their actions. Seeing their choices as they perceive them, in turn, entails a focus on options and reasons rather than distant causes, determinism or (in the case of Plum) manipulation. However, taking up an agential perspective on others does not *entail* holding them responsible or blaming them when they do wrong; I leave it open that there might be reasons independent of *the four-case manipulation argument* not to hold people responsible for what they do. The agential perspective does, however, extinguish our non-responsibility intuitions about classic manipulation cases. If we ought to regard Plum in this way, as I have argued, the four-case manipulation argument does not get off the ground. I leave it open that Pereboom's hard incompatibilism might be true after all; my business in this paper is merely to undermine one of his primary arguments for it.

Now, Pereboom might agree with quite a lot of what I have argued for so far. He does argue that if the world is deterministic, it can still be rational to consider various options and deliberate about what to do (Pereboom 2008, 2014: Chap. 5). Possibly, Pereboom agrees as well that it makes sense for Plum to disregard the fact that he is manipulated when deliberating about what to do, and for us to disregard it in much of our interactions with him. Still, according to Pereboom, in order to determine whether Plum can really be *morally responsible* for what he does, we must adopt a more detached, causal perspective and focus on the causal chain behind his actions. It is only when we do this that trustworthy intuitions arise. However, this claim clearly needs to be argued for, and I have already pointed out the problems with the idea that our intuitions become more trustworthy the more information we try to simultaneously focus on. First, it is not obviously true that the

more information we try to focus on the better, since irrelevant information might disturb our intuitions rather than help them. Second, regarding an agent and his choice from an agential perspective, putting oneself in his shoes, means regarding him *differently*—it is not just a case of arbitrarily ignoring certain pieces of information while keeping others.

It might still be argued that we cannot account for certain intuitively plausible cases of diminished responsibility, or the tracing of responsibility back to previous actions, without leaving the agential perspective and focus on the causal chains leading up to the agent's deliberation and decision. This argument, however, can be resisted.

6 Diminished responsibility and tracing

Fischer and Ravizza (1998: 49–50, Chap. 7–8) argues that moral responsibility judgments must take into account the agent's history, and thus cannot be based solely on the agent's options, reasons and choice at the time. Some of their arguments rely on manipulation cases, and can be left aside since I have discussed manipulation already. However, they also write that an agent might intuitively not be morally responsible for his current actions because he was abused as a child (ibid: 187). In a similar vein, Strawson writes that when we see a wrongdoer as "peculiarly unfortunate in his formative circumstances", our resentment tends to recede (Strawson 1962/2013: 70). Fischer and Ravizza furthermore writes that a theory of moral responsibility must allow for *tracing*; a drunk driver currently not reasons-responsive can still be held morally responsible for running over someone, if his current action was caused by a previous action (voluntarily drinking large quantities of alcohol) for which he *was* responsible.

Starting with wrongdoers from harsh backgrounds, it is easy to see what the prima facie problem is for my account. Just like Plum thinking "whatever I do I will be manipulated into doing" cannot help him make up his mind, the person from a harsh background cannot decide what to do by focusing on the fact that his actions might be caused by his childhood. Just like the Hologram Doctor occasionally pondered his manipulation, our poor wrongdoer might think about his harsh background from time to time; perhaps he even does so a lot. Still, when deciding whether to, e.g., steal an object, he has to shift focus to his reasons for and against doing so—on the one hand he might get caught, on the other hand the thing is worth a lot of money, and so on. If we are to judge whether he is morally responsible for the theft by putting ourselves into his shoes when he committed it, the more distant causes of his action will fall out of view.

Now, a proponent of the causal perspective as the correct one from which to make moral responsibility judgments might argue in the following way: The thief's harsh background is intuitively relevant for his moral responsibility. Therefore, the causal chain leading up to an agent's deliberation and decision *is* relevant for moral responsibility after all; we cannot make proper moral responsibility judgments unless we leave the agential perspective, take a step back and investigate what caused the agent to become the kind of person that he is and do what he does.

However, the importance of not being too harsh on people in our judgments might actually be *better* taken into account if we go the opposite route. Instead of looking at the offender from a detached, causal perspective, let us really step into his shoes before judging him, and take his *difficulties* fully into account. If he suffers from, e.g., terrible impulse control, feels pressured by his peers and so on, making his choice a hard one, these factors can be fully appreciated when seeing his choice the way he does. Dana Nelkin and Susan Wolf write that it is part of common-sense morality that an agent is less blameworthy for doing wrong when it is very difficult for her to do right, and that a good theory of moral responsibility must take this into account (Nelkin 2016; Wolf 1990: 86–87). It is not problematic or counter intuitive to claim that a certain history of the agent might play an important *epistemic* role and provide evidence that the agent suffers from certain difficulties, but it is the difficulties themselves that do the responsibility diminishing job. If those difficulties are severe enough, they might inhibit responsibility completely and provide a full excuse.

It is, in fact, the suggestion that causal history is relevant *in itself* that has counter intuitive implications. McKenna gives us the example of Ann, whose life values (life is precious, she must live it to the fullest and so on) are shaped at an early age by factors completely outside of her control, namely the death in cancer of a parent (McKenna 2008: 156). When Ann acts from those values, she judges herself morally responsible for doing so, and this is surely the intuitive, common-sense judgment. This is unsurprising on Nelkin and Wolf's view, according to which difficulties in doing right mitigates moral responsibility for wrong-doing, because Ann's causal history did not cause her to experience such difficulties. On the suggestion that it is responsibility undermining in itself when someone is caused to do what she does by causes beyond her control, however, we have to draw the counter-intuitive conclusion that Ann is not morally responsible for what she does, when acting on her deepest values.

Finally, let us take a look at tracing. Although the agent obviously knows about her earlier choices and might think about them a lot, when finally making a decision, she must focus forward on her options and reasons, rather than pondering how she got there. It thus seems like taking up the agential perspective and seeing the situation the way the agent does at the moment of choice precludes the possibility of tracing her responsibility back to earlier choices. However, tracing is not obviously needed in a moral responsibility theory.

The stock example of indirect moral responsibility, used by Fischer and Ravizza as well as countless other philosophers, is the drunk driver, who runs someone over while lacking reasons-responsiveness and other standard moral responsibility requirements due to his serious intoxication. It is debatable how plausible this example is as it is normally presented (can you really utterly lack reasons-responsiveness and other standard moral responsibility requirements while still being capable of driving a car, however badly? Khoury 2012: 193). Still, if one has doubts about stock examples, it is easy enough to tweak them in order to make it more plausible that the driver lacked moral responsibility relevant capacities at the time of the kill—we could, for instance, replace alcohol with a drug that, after the first high, causes you to fall asleep. The agent takes the drug at a party, thinks he

will make it home before sleep hits him, but alas, falls asleep at the wheel and kills someone (Khoury 2012: 194). I believe with Khoury (2012) and King (2011) that such cases can be handled without appealing to tracing. We cannot put ourselves in the agent's shoes at the time of the kill, since he is asleep then, but we can put ourselves in his shoes when he takes the pill. We can assess his options (taking or not taking the pill), his reasons for each option (having a good time, not putting others in danger), and blame him for making a horrible choice.

Khoury (2012) argues that it is really only the choice one is responsible for, although the consequences (in this scenario, a dead pedestrian) can serve as evidence that the choice really was a horrible one. King (2011), on the other hand, argues for including consequences when judging what the agent was responsible for (at least insofar as they were foreseeable), but denies that we need a special moral responsibility mechanism for including them. Imagine three soldiers fleeing from and trying to stop pursuing enemies. The first soldier throws a hand grenade back at them, so that they are killed a few seconds after he threw it. The second one sets up a land mine to explode when the pursuing soldiers, after a few minutes, step on it. The third one sets up explosives that do not go off, killing the enemies, until several hours later. We do not need tracing in the latter cases merely because more time has passed between the soldier setting up the mine or the explosives and the enemies dying; the responsibility is still direct. King further argues that examples like that of drunk or drugged driving are close enough to the above that we do not need any particular explanation for those either. Either the driver intentionally exposed other people to a risk that ended up getting them killed (just like the soldier setting up mines, knowing full well that there is a risk an innocent might get killed instead of his enemies), or he recklessly did so—but there is no need to assume a different kind of moral responsibility than the ordinary one.

The debate about tracing in moral responsibility is a large one, and I cannot resolve it within the confines of this paper. But as long as there are good arguments to the effect that we can do without tracing, the fact that we cannot trace moral responsibility while sticking to an agential perspective when making moral responsibility judgments is no decisive objection to my account.

7 Conclusion

So where does all this leave us? The hard-liner must accept that Plum 1 and 2 are morally responsible for what they do, and when reading Pereboom's descriptions of the cases, this is counter-intuitive. Pereboom, on the other hand, must accept that fictional characters like Star Trek Voyager's Hologram Doctor lack moral responsibility, and when partaking of those stories, *this* is counter-intuitive. So far, we have a genuine stalemate where both sides need to accept a counter-intuitive conclusion. Furthermore, a hard-liner who is committed to the idea that the agent's action being caused by factors beyond her control is *not* in itself responsibility undermining, only resulting difficulties are, can make intuitively plausible judgments about both the criminal from a harsh background and McKenna's Ann, whereas the incompatibilist who argues that such causation *is* responsibility

undermining in itself must make the counter-intuitive judgment that Ann's history undermines her responsibility for important decisions based on her deepest values. The hard-liner who, like me, advocates making moral responsibility judgments from an agential perspective where we try to put ourselves in the agent's shoes, can point to the fact that this perspective on other people is both the natural and the morally obligatory one for ordinary, everyday interaction. The incompatibilist who wants to claim that we ought to make moral responsibility judgments from a detached causal perspective therefore owes us an argument as to why that is. As long as no such argument is presented, we do not have a genuine stalemate here—rather, the hard-liner is slightly ahead of the incompatibilist in the race.

Acknowledgements Special thanks to Gunnar Björnsson and Ben Matheson for detailed comments on earlier drafts. Versions of this paper or the central argument in it have also been presented at the Manipulation and Moral Responsibility workshop at the University of Edinburgh 2016, Sense of Agency workshop at Université catholique de Louvain in 2017, and at the 2017 workshop with Dana Nelkin at the University of Gothenburg, by the Gothenburg Responsibility Project. This work was supported by Vetenskapsrådet, Grant No. 2014-40.

References

Baker, L. (2006). Moral responsibility without libertarianism. *Noûs, 40*(2), 307–330.
Bok, H. (1998). *Freedom and responsibility*. Princeton: Princeton University Press.
Demetriou, K. (2010). The soft-line solution to Pereboom's four-case argument. *Australasian Journal of Philosophy, 88*(4), 595–617.
Fischer, J. M. (2004). Responsibility and manipulation. *The Journal of Ethics, 8*(2), 145–177.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5–20.
Hume, D. (1739/1978). *A treatise of human nature*. Oxford: Oxford University Press.
Jeppsson, S. (2016). Reasons, determinism and the ability to do otherwise. *Ethical Theory and Moral Practice, 19*(5), 1225–1240.
Kapitan, T. (1986). Deliberation and the presumption of open alternatives. *Philosophical Quarterly, 36*(143), 230–251.
Khoury, A. C. (2012). Responsibility, tracing and consequences. *Canadian Journal of Philosophy, 42*(3–4), 187–207.
King, M. (2011). Traction without tracing: A (partial) solution for control-based accounts of moral responsibility. *European Journal of Philosophy., 22*(3), 463–482.
Korsgaard, C. M. (1996a). *Creating the kingdom of ends*. Cambridge: Cambridge University Press.
Korsgaard, C. M. (1996b). *The sources of normativity*. Cambridge: Cambridge University Press.
Matheson, B. (2014). Compatibilism and personal identity. *Philosophical Studies, 170*(2), 317–334.
Matheson, B. (2016). In defence of the four-case argument. *Philosophical Studies, 173*(7), 1963–1982.
Matheson, B. (forthcoming). Towards a structural ownership condition on moral responsibility. *Canadian Journal of Philosophy,* 1–23. https://doi.org/10.1080/00455091.2018.1480853.
McKenna, M. (2005). The relationship between autonomous and morally responsible agency. In J. S. Taylor (Ed.), *Personal autonomy* (pp. 205–234). Cambridge: Cambridge University press.

McKenna, M. (2008). A hard-line reply to Pereboom's four-case manipulation argument. *Philosophy and Phenomenological Research, 77*(1), 142–159.

McKenna, M. (2013). Resisting the manipulation argument: A hard-liner takes it on the chin. *Philosophy and Phenomenological Research, 89*(2), 467–484.

Mele, A. (2005). A critique of Pereboom's 'four-case' argument for incompatibilism. *Analysis, 65*(1), 75–80.

Milam, P.-E. (2016). Reactive attitudes and personal relationships. *Canadian Journal of Philosophy, 46*(1), 102–122.

Nelkin, D. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs, 50*(2), 356–378.

Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.

Pereboom, D. (2008). A compatibilist account of the epistemic conditions of rational deliberation. *The Journal of Ethics, 12*(3–4), 287–306.

Pereboom, D. (2014). *Free will, agency and meaning in life*. Oxford: Oxford University Press.

Pereoom, D. (2005). Defending hard incompatibilism. *Midwest Studies in Philosophy, 29*(1), 275–286.

Russell, P. (2010). Selective hard compatibilism. In J. Campbell, M. O'Rourke & H. Silverstein (Eds.), *Action, ethics and responsibility: Topics in contemporary philosophy* (Vol. 7, pp. 149–173). Cambridge, Mass: MIT Press.

Sommers, T. (2007). The objective attitude. *The Philosophical Quarterly., 57*(228), 321–341.

Strawson, P. (1962). Freedom and resentment. In *Proceedings of the British Academy, 48*,1–25. (Reprinted in *The philosophy of free will. Essential readings from the contemporary debate*, pp. 63–83, by P. Russell, & O. Deery, Eds., 2013, New York: Oxford University Press. Pagination refers to this reprint).

Uhlmann, E. L., et al. (2009). The motivated use of moral principles. *Judgment and Decision Making., 4*(6), 476–491.

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. Oxford: Oxford University Press.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge: Harvard University Press.

Wolf, S. (1990). *Freedom within reason*. Oxford: Oxford University Press.