*Canadian Journal of Philosophy (2025), 55: 1, 84–98*
doi:10.1017/can.2025.10044

ARTICLE

# Masked Abilities and the Four-Case Manipulation Argument

Gus Turyn

Philosophy, Cornell University, Ithaca, New York, USA
Email: mat298@cornell.edu

**Abstract**

I offer a novel dispositional reply to Derk Pereboom's four-case manipulation argument. Drawing on recent work in the metaphysics of dispositions, I argue that manipulated agents' rational abilities are masked—prevented from manifesting as they otherwise would—by neuroscientists' manipulation. I argue that masking better explains why manipulated agents are not responsible for their actions than causal determinism does, as we ordinarily take masks to explain why agents are not morally responsible for their actions or inaction. Because causal determinism is not a mask, there is a relevant difference between manipulation and causal determinism, and the four-case argument fails.

**Keywords:** manipulation arguments; free will; moral responsibility; dispositions; masked dispositions

## 1. Introduction

Manipulation arguments present a significant threat to compatibilism about free will and causal determinism. Advocates of manipulation arguments hold that there is no difference between agents who are manipulated by others and agents who are causally determined. Because manipulated agents are not responsible for their actions, they argue, causally determined agents are likewise not responsible.

Derk Pereboom (1995, 2001, 2014, 2020, 2022) has advanced an influential manipulation argument proceeding through four cases describing Plum, an agent who is considering whether or not to kill White.[1] In the first, Plum1 is actively manipulated by neuroscientists who alter his neural states to make him reason more egoistically than he otherwise would, resulting in him choosing to kill White. In the second, Plum2 is programmed at birth by neuroscientists who design him such that he will often reason egoistically; when presented with the opportunity to kill White, he chooses to do so. In the third, Plum3 is raised in a community that encourages egoistic reasoning and chooses to kill White because of this influence. In the fourth, Plum4 is an ordinary causally determined agent who is not raised in an especially egoistic community, but who nonetheless is causally determined to choose to kill White. Pereboom argues that because there is no morally relevant difference between these agents, and because Plum1 is not morally responsible for his actions, Plum4 is not morally responsible for his actions.

[1]My focus in this paper is on Pereboom's four-case manipulation argument, but see Mele (2006), Cyr (2016), and Matheson (2016) for other manipulation arguments. See Shabo (2010) and Todd (2011, 2016) for further defenses of manipulation arguments.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Canadian Journal of Philosophy, Inc. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (http://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

My aim here is to offer a novel dispositional reply to Pereboom's manipulation argument. Dispositional compatibilists (Fara, 2008; Smith, 2003; Vihvelin, 2004, 2013) hold that having free will is a matter of having abilities that are analyzable as dispositions. Because the existence of unmanifested dispositions such as fragility is not threatened by the truth of determinism, they hold, the existence of unexercised abilities—including the ability to do otherwise—is compatible with the truth of determinism. Dispositionalists have thus far largely ignored manipulation arguments, focusing their energies instead on van Inwagen's (1983) Consequence Argument and Frankfurt's (1969) challenge to the Principle of Alternative Possibilities. I will address this gap in the literature by arguing that when coupled with an adequate understanding of *masks*, a dispositional account of free will has the resources to overcome the challenge presented by Pereboom's argument.

A disposition is masked when something would interfere with the disposition's manifestation without altering the disposition.[2] Johnston (1992) offers the example of a glass supported by a piece of styrofoam: though the glass is fragile, the stimulus (being dropped or struck) would not bring about the manifestation (breaking).[3] I will argue that the manipulators in Pereboom's first case occupy the same role: though manipulated agents retain their abilities to deliberate on the basis of reasons and make decisions, those abilities are masked. While Pereboom argues that the best explanation for why manipulated agents are not morally responsible is that they are causally determined by events beyond their control, I will argue that their decision-making abilities' being masked better explains why they are not responsible. Agents in causally determined universes who we ordinarily assume act freely, however, do not typically have their decision-making abilities masked. As such, a principled distinction can be drawn between manipulation and causal determinism, affording the dispositional compatibilist a means of distinguishing Pereboom's first and fourth cases.[4]

If my reply succeeds, the dispositional account is an especially attractive option for compatibilists. While what dispositionalists say about the Consequence Argument and Frankfurt-style cases is controversial, any plausibility that the view has stems from its capacity to secure compatibilists the ability to do otherwise in a deterministic world. Though dispositionalism grounds our freedom in our ability to do otherwise and is thus a *leeway* account, it can also provide the causal condition offered by *sourcehood* compatibilists necessary to undercut the manipulation argument.[5] Many take extant unaddressed challenges to the view to undercut its plausibility, but if my arguments succeed, the view warrants further consideration as a path forward for compatibilists.[6]

This paper proceeds as follows. In Section 2, I will discuss dispositional compatibilism and the connections between abilities and dispositions in further detail. In Section 3, I will discuss masked dispositions and abilities in further detail. In Section 4, I will argue that certain features of masking allow dispositionalists to draw a morally relevant distinction between the first and second cases in Pereboom's argument, allowing for a principled rejection of the argument. I will address potential objections in Section 5 and the dialectical role of my argument in Section 6.

[2]This is only a rough gloss on what masks are; I will discuss masks in further detail in Section 3.

[3]See Bird (1998) for another early discussion of masks.

[4]The reply that I will develop falls under the general category of 'soft-line' replies (to borrow McKenna's (2008) terminology). Soft-liners hold that there is a metaphysical difference between manipulation and determinism, while hard-liners deny that manipulated agents are not responsible for their actions. I will discuss precisely where my account falls between these camps in Section 6.

[5]See Frankfurt (1969, 1971) and Sartorio (2016, 2023) for sourcehood accounts.

[6]Clarke's (2009) challenge is perhaps the most significant, but see Whittle (2010), Janzen (2016), and Vetter (2019).

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

## 2. Dispositions and Abilities

Ordinary objects and agents alike possess a host of ordinary dispositions. Fragile glasses are disposed to break when dropped; loquacious agents are disposed to talk frequently. These dispositional properties are modal properties. Whether a glass is fragile or an agent is loquacious is not a matter of the actual sequence of events, but is instead a matter of how objects and agents *would* (or *might*) behave under a range of possible circumstances.[7] A glass that never breaks is still fragile; a loquacious agent remains loquacious even when they are sitting silently in a lecture. Further, as noted above, the truth of determinism is irrelevant to the question of whether a given glass is fragile: if it is entailed by the past and the laws that a particular glass never breaks, that glass is still fragile.

Agential abilities are also modal properties. I am able to walk exactly 612 steps forward, then turn around and walk exactly 611 steps back. I have never done so (and imagine that I will never do so); my possession of this ability is a matter of what I *would* (or *might*) do under a range of possible circumstances. The "new dispositionalism," as Clarke (2009) labeled it, is a family of compatibilist theories of free will that center around the observation that dispositions and abilities are alike in this regard. Its primary proponents—Michael Smith (2003), Kadri Vihvelin (2004, 2013), and Michael Fara (2008)—each offer different views of the precise connection between dispositions and free will. For Smith (2003, 27–30), our abilities to *A* are underwritten by our rational capacities to form beliefs and desires; for Vihvelin (2013, 187), to have an ability to *A* just is to have a particular disposition to *A* as the result of trying; for Fara (2008, 848–849), to have an ability to *A* depends on having a disposition to succeed at *A*-ing when one tries to *A*.[8] For the new dispositionalists, that abilities bear some such relationship to dispositions means that the existence of unexercised abilities, too, is compatible with the truth of determinism.

Classical compatibilists (e.g., Ayer, 1954) argued that an agent has free will just in case she would have done otherwise had she chosen to do otherwise. On such a view, determinism does not rule out free will: in many cases, causally determined agents would have done otherwise had they chosen to do so. The problems with such a view, of course, are closely related to the problems with counterfactual analyses of dispositions. Counterfactual analyses of dispositions are famously vulnerable to two kinds of counterexamples: *masks* and *finks*.[9] A disposition is finkish when the same conditions that would stimulate the disposition would alter the disposition; Lewis (1997) gives the memorable example of a sorcerer who would alter a glass' molecular structure if the glass were ever struck.[10] While the glass is fragile, it would cease to be fragile if the conditions that would stimulate its fragility ever came about. A disposition is masked (roughly speaking) when something would prevent the disposition from manifesting as it normally does while leaving the disposition itself intact. As noted above, while a glass supported by a piece of styrofoam is fragile, the conditions that would stimulate its fragility would not cause it to break.

If abilities are reducible to dispositions, then they too can be finkish and masked. We can imagine the same sorcerer preparing to alter a person's musculature if that person were to try to lift a weight; while the person has the ability to lift the weight, they would lose the ability if they attempted to exercise it. We can also imagine the same person handcuffed; though they have the ability, they are unable to exercise it due to their being handcuffed.

[7]Of course, dispositions have categorical bases: glasses are fragile in virtue of the fact that they have particular (nonmodal) chemical structures; agents are loquacious in virtue of the fact that they have particular (nonmodal) neural structures.

[8]Can these views account for our abilities to try? Arguably not: Fara (2008, 849) addresses this point and notes that one cannot in principle try to try. See Clarke (2009) for further discussion of dispositionalism's trouble here.

[9]Both masks and finks were originally raised as counterexamples to the 'simple conditional analysis of dispositions' which Lewis puts as follows:

> Something *x* is disposed at time *t* to give response *r* to stimulus *s* iff, if *x* were to undergo stimulus *s* at time *t*, *x* would give response *r* (Lewis, 1997, 143).

[10]The problem of finks was first raised in Martin (1994).

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

An enormous literature has emerged since the problems of finks and masks were first raised that grapples with how analyses of dispositions might avoid these problems. While I will remain neutral here about what the right account of dispositions is, if there is some account that can overcome these problems, then the problems of masks and finks are also tractable for dispositional accounts of abilities.[11]

## 3. Masks

Masks are ordinarily construed as phenomena that would prevent a disposition from manifesting while leaving the disposition itself intact. Fragile glasses remain fragile even when wrapped in bubble wrap. As noted above, agents' abilities can also be masked: agents whose hands are tied behind their backs do not lose the ability to shake hands with strangers.

Masks have been the subject of a good deal of controversy in the metaphysics of dispositions.[12] While I will not assume any account or another of what masks are here, my argument in what follows depends on the truth of two principles about masking.

*Principle 1*: Masks come in degrees.

Elsewhere, I (Turyn, 2021) have argued that masks can come in degrees and, like dispositions themselves, are gradable along two dimensions.[13] Suppose that one glass is wrapped in a thin sheet of bubble wrap and another is wrapped in a thick sheet of bubble wrap. Suppose that both glasses are dropped, that neither breaks, and that each sheet of bubble wrap plays a causal role in the respective glasses not breaking. Any account of masks must grant that both sheets of bubble wrap are masks. Yet it is clearly better to wrap one's favorite glass in a thicker sheet of bubble wrap; while both sheets of bubble wrap *can* prevent the glass from breaking, the thicker sheet is likelier to do so.

Similarly, suppose that both glasses are dropped and that both break slightly upon being dropped —the glass in the thin sheet suffers a large crack while the glass in the thick sheet suffers only a small chip. Though both glasses manifest fragility as the result of being dropped, both sheets of bubble wrap decrease the degree to which they would otherwise do so. While not all extant accounts of masks are sensitive to the fact that masks are gradable, any adequate account of masking should account for such cases: it would be odd to deny that either of these are cases of masking.[14,15]

*Principle 2*: The mere presence of a mask does not entail that a disposition's manifestation does not appear.

---

[11]Among the new dispositionalists, Vihvelin (2013) has done the most to address the problems of masks and finks. She adopts Lewis's (1997) solution to the problem of finks and a modified version of Manley and Wasserman's (2008) solution to the problem of masks.

[12]See e.g. Lewis (1997), Manley and Wasserman (2008), Choi (2011), and Turyn (2021) for discussions of the various controversies around masks.

[13]See Manley and Wasserman (2007, 2008) for discussion of the gradability of dispositions. Manley and Wasserman (2008) argue that dispositions are gradable along the same two dimensions that I will discuss here: one glass can be more fragile than another because it is *likelier* to break and one glass can be more fragile than another because it would break *to a greater degree* if dropped or struck.

[14]Indeed, while most accounts of masks (e.g. Fara, 2008; Gebharter and Fischer, 2021) are silent on such cases, they could easily be adapted to account for them.

[15]To date, the only account of masks of which I am aware that acknowledges that masks come in degrees is my own (Turyn, 2021). One might worry that my account is too permissive with respect to what counts as a mask, as I hold that *anything* that decreases the degree to which a disposition would otherwise manifest counts as a mask. This might lead to a problematic overabundance of masks: anything, even a small smudge on an otherwise perfectly functioning lightbulb, might thus count as a mask. (Thanks to an anonymous referee for raising this objection.) What I will argue in what follows does not depend on this account of masks, so I will not address this worry in full here, but note that my account could easily be modified to restrict masks to phenomena that, say, reduce the degree to which a disposition manifests *to a sufficiently high degree*, where context determines how high the degree must be for something to count as a mask. In certain scientific or artistic contexts in which the amount of light is quite important, a smudge might indeed count as a mask; in other contexts, the threshold might be higher.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

On the assumption that Principle 1 is true, Principle 2 is also true: a glass wrapped in bubble wrap might still manifest its fragility, but only to a lesser degree than it would if the glass were not wrapped in bubble wrap if, for instance, it chips when it is dropped rather than shattering entirely. But there are other cases in which a disposition's characteristic manifestation might appear despite the presence of a mask that are relevant to the arguments that I will advance in what follows.

Masks are one of several widely-discussed kinds of dispositional *interferers*: phenomena that affect whether a disposition manifests as it normally would. Beyond masks and finks, *mimics* also present *prima facie* trouble to conditional analyses of dispositions. A mimic is something that brings about the characteristic stimulus of a disposition through some process other than the causal sequence that ordinarily leads from the disposition's stimulus to its manifestation. When a brick is dropped onto an explosive device primed to go off as soon as it makes firm contact with something, the brick will break; though it is not fragile, its being dropped would lead to its breaking. Lewis (1997, 153) (characteristically) offers the more memorable example of the "Hater of Styrofoam," an agent who cannot stand the sound that styrofoam makes when it drops and so destroys pieces of styrofoam whenever he hears them strike the floor. Though any given piece of styrofoam in the Hater of Styrofoam's vicinity is not fragile, it would break when dropped because the Hater of Styrofoam would destroy it.

Masks and mimics can co-occur.[16] Suppose that the Hater of Styrofoam heard a glass protected in a piece of styrofoam packaging fall to the ground. While the styrofoam packaging masks the glass' fragility, the Hater of Styrofoam simultaneously mimics it by stomping on the packaging, destroying both the styrofoam and the glass. (We can imagine that he is exceptionally well-coordinated and stomps on it at the precise moment that it strikes the ground.) Although the glass did break when dropped, it did not do so in virtue of the causal sequence that would ordinarily result in its doing so; it does so because an external influence interfered with it.

Deviant causal chains such as this can, of course, also result in the characteristic manifestations of agents' abilities appearing. Suppose that Adam is a skilled golfer preparing to putt a golf ball. Right after he putts, an unexpected gust of wind knocks the ball slightly off of its path. Fortunately, a sprinkler happens to go off an instant later; a jet of water knocks the ball back onto its path and into the hole. Though Adam is able to sink the putt and though his trying to sink it led to his actually sinking it, his ability was masked and then mimicked. It is because of these simultaneous interferences that the ball makes it into the hole.

### 3.1. Masked abilities and blameworthiness

I will spend the remainder of this section arguing that masks sometimes undermine or mitigate blameworthiness. Given some other background conditions, an agent's ability to make decisions and act on the basis of reasons remaining unmasked is a necessary condition on their being blameworthy. This condition is no more or less important than other necessary conditions on blameworthiness, such as the satisfaction of a particular epistemic condition.

Reflections on and intuitions about our practices will help motivate an anti-masking condition on blameworthiness. Consider some familiar cases from Strawson (1962). Anne is pushed and accidentally steps on Ben's foot. While Ben might initially blame her for the harm she caused, he would typically cease to do so upon discovering that she had been pushed. Zander, who has been significantly stressed out over his work recently, rudely snaps at Yuri. While Yuri might initially blame him for doing so, he would typically cease to do so upon reminding himself how stressed out Zander has been. Though Strawson (unsurprisingly) did not import metaphysical terminology into cases such as this, doing so is easy enough: Anne's ability to control her bodily movements is masked

[16]See Turyn (2021, section 3) for discussion of similar cases.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

by her being pushed and Zander's ability to judge what he should say to others is masked by his stress level.[17]

Other everyday cases also motivate a connection between masking and blameworthiness. Suppose that Doug is a paramedic who sees another agent suffering from respiratory failure. Though Doug is able to insert a respiratory device and save the other agent's life, he is currently handcuffed to a streetlight. Though he struggles against the handcuffs, he fails to break free in time to reach the other agent. While Doug might be blameworthy for failing to exercise his ability to save the other agent's life under other circumstances, his ability is masked here.

Such cases motivate an *anti-masking* condition on blameworthiness because these are the precise considerations to which we appeal when explaining why such agents are not blameworthy. My proposal is that one necessary condition on an agent's being blameworthy for some wrong action is that the abilities in virtue of which the agent who performed that action is free and responsible—the abilities to decide and act on the basis of reasons, the ability to do otherwise, among others—are not masked at the time at which the agent deliberates and acts.

One might immediately object to this proposal on the grounds that we are often blameworthy for our actions even when our abilities are masked. I will now argue that these challenges do not ultimately undermine the motivation for an anti-masking condition on blameworthiness. (For ease of discussion, I will begin by considering cases involving masks that *completely* prevent a disposition from manifesting, then introduce complications raised by degrees of masking.)

First, to the extent that the new dispositionalism is an attractive view, it is because it has the resources to secure the compatibility of free will and determinism while holding that the Principle of Alternative Possibilities is true. Smith (2003) and Fara (2008) both argue that in Frankfurt-style cases, the counterfactual intervener serves as a mask,[18] but that the would-be manipulated agent remains able to do otherwise and is morally responsible for his actions.[19] If we held that one's abilities being masked rendered one non-responsible, then the new dispositionalists would be forced to reject this analysis of Frankfurt-style cases.

We can distinguish between such cases, however, by delineating cases in which masks play a causal role in the actual sequence of events that leads to the lack of a manifestation of an agent's ability from those in which masks do not. Anne's ability to control her bodily movements is causally interfered with by the push and Doug's ability to save the other agent's life is causally interfered with by the handcuffs; the counterfactual intervener in a Frankfurt-style case plays no such causal role in Jones's action.

Reflection on these cases suggests that we should revise the anti-masking condition to focus on cases in which masks *actually interfere* with our relevant abilities and dispositions. But such a condition by itself is also subject to counterexamples. Suppose, for instance, that Doug had handcuffed himself to the streetlight to prevent himself from saving the other agent's life (perhaps because he wanted the other agent to die but knew that his life-saving instincts would kick in as soon as he saw the other agent enter respiratory failure). In such a case, Doug would intuitively still be blameworthy for not saving the other agent's life. Yet if Doug's life-saving instincts kicked in and he attempted to free himself to save the other agent, the handcuffs would indeed play a causal role in the sequence of events that led to his ability failing to manifest.

[17]One might object that Anne temporarily loses her ability to control her body when she is pushed. She could, after all, make a plea that she could not have stopped herself from stepping on Ben's foot and it would seem counterintuitive to say that this plea is false. But the standards by which we evaluate modal claims in ordinary discourse clearly differ from the standards by which we evaluate modal claims in discussions of free will and responsibility. Anne's plea is true in ordinary contexts such as this one but not true, I hold, in contexts that demand higher standards for modal claims.

[18]Strictly speaking, Fara (2008) holds that masks are a matter of the actual sequence of events, not a counterfactual matter, so Jones's abilities are not masked but instead *would be* masked if he were to try to do otherwise. Unlike Fara, though, I think that a disposition can be masked even if it is not stimulated, so I will speak here in terms of Jones's ability being

[19]Vihvelin (2004, 2013) holds that such interveners sometimes work as finks and sometimes as masks, depending on how the details of the case are spelled out.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

We can account for such cases by adding a historical component to the anti-masking condition.[20] A mask undermines blameworthiness only if the agent is not themself responsible for the presence of that mask (i.e., they neither voluntarily brought about that their ability is masked nor did they negligently allow it to come about). Such a condition will allow us to delineate cases in which a mask that plays a causal role in an ability not manifesting undermines blameworthiness from those in which it does not; if Doug was involuntarily handcuffed to the streetlight *and* if the handcuffs play a causal role in Doug's failing to save the other agent, he is not blameworthy; if Doug voluntarily brought it about that he was handcuffed, then regardless of whether the handcuffs play a causal role in his failing to save the other agent, he is blameworthy. (I will say more about the kind of historical considerations relevant to my argument in Section 5 when I consider two objections to my argument.)

We thus have an anti-masking condition on the responsible agency that can explain why agents whose abilities are actually interfered with by the presence of a mask are not blameworthy for their actions. I want to turn now to degrees of masking. Just as all-out masking entirely undermines an agent's blameworthiness, I will argue in what follows that degrees of masking can do so as well.

Suppose that Eric is pushed firmly enough that he loses his balance, but not so hard that he cannot partially exercise his ability to control his bodily movements. Suppose that he realizes that he can either direct his body such that he falls into a pile of garbage, which would dirty his outfit, or towards Fabian, which would result in him stomping on Fabian's foot. Eric's ability to control his bodily movements is indeed masked, but to a lesser degree than Anne's is. Should Eric exercise his ability to control his bodily movements in such a way that he stomps on Fabian's foot, he is intuitively more blameworthy for his action than Anne is for hers, but still intuitively less blameworthy for his action than an agent without a masked ability would be for theirs.

Similarly, consider an example from Coates and Swenson (2013). Suppose that Marcia breaks her promise to pick you up at the airport, but explains that she has been suffering from depression recently.[21] On the plausible assumption that depression can render us less likely to recognize reasons, make decisions, and execute our intentions, Marcia's ability to pick you up at the airport is masked by her depression (to the extent that she is rendered less likely to pick you up in virtue of the fact that she suffers from depression). And it seems that her abilities being masked explains her being less blameworthy than another agent would be for the same action if that other agent did not suffer from depression.

The initial motivation for an anti-masking condition on responsible agency is that we often appeal to the presence of masks to explain why agents are not blameworthy for certain actions. With refinements to avoid apparent counterexamples, a history-sensitive anti-masking condition can distinguish between cases in which the presence of a mask undermines or mitigates blameworthiness from those in which it does not. And degrees of masking can explain reduced degrees of blameworthiness and comparative judgments of blameworthiness across different agents. Just as agents are sometimes less blameworthy for their actions in virtue of their failing to satisfy an epistemic condition, agents are sometimes less blameworthy for their actions in virtue of their abilities being masked.

[20]This is not the only theoretical option available here; note that we could instead restrict the anti-masking condition to direct or non-derivative blameworthiness. Thanks to an anonymous referee for noting this theoretical option.

[21]Some (e.g., Choi, 2017) hold that the intrinsic properties of an entity cannot mask that entity's dispositions. Even if one grants that intrinsic masks are possible, however, one might object that depression robs us of our abilities rather than masks those abilities. If either of these claims are correct, then this case does not illustrate the relevant point. But sufficiently similar cases can be constructed that illustrate the general point I want to make here. Suppose that an annoyed neighbor constructs an elaborate puzzle blocking Marcia's driveway. Marcia could solve the puzzle if she tried sufficiently hard, but is unlikely to do so before she would need to leave for the airport. Her ability in this case is uncontroversially masked, though not *entirely* masked.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

## 4. Masked Abilities and Manipulation

We can now turn to the four-case manipulation argument. As discussed above, Pereboom argues that a manipulated agent controlled by malicious neuroscientists can satisfy any and all proposed compatibilist conditions for free will and yet still not be morally responsible. Ultimately, Pereboom argues, the best explanation for why manipulated agents who satisfy all compatibilist conditions for free will are not morally responsible is because their decisions and actions are causally determined. Further, because there is no difference relevant to moral responsibility between a manipulated agent and a causally determined agent, Pereboom argues, causally determined agents are also not free. Pereboom's argument relies on two premises:

(1) There is no difference relevant to moral responsibility between Plum1, Plum2, Plum3, and Plum4.
(2) Plum1 is not morally responsible for killing White.

If both premises are true, then Plum4 is not morally responsible for his actions, a conclusion that can be generalized to all causally determined agents. Compatibilists who deny the first premise—the *No Difference* premise—offer soft-line replies; compatibilists who deny the second—the *No Responsibility* premise—offer hard-line replies.[22]

What makes the four-case argument so powerful—more powerful, in my view, than any other manipulation argument—is the strength of the intuition that Plum1 is not responsible for his actions. An agent who is actively manipulated by others far more plausibly lacks responsibility for their actions than does an agent who happens to be designed in such a way that they commit some bad action. Further, because we actually encounter actively manipulated agents in the real world, it is plausible that these intuitions are more reliable than our intuitions about designed agents are.

My aim here is to argue that the upshot of the previous section is that dispositionalists have the means to show that the *No Difference* premise is false. The best explanation for why Plum1 is not morally responsible for his actions is that his abilities are masked by the neuroscientists' manipulation, not that his actions are causally determined. Because Plum4's abilities are in no way masked, there is a difference relevant to moral responsibility between cases of manipulation and ordinary cases, and the manipulation argument fails. As noted above, my focus in what follows will be on Plum1's ability to deliberate and act on the basis of reasons, an ability which Vihvelin (2013, 189) argues is necessary for being a morally responsible agent.

Vihvelin (2013, 148–155) is the only new dispositionalist to address the manipulation argument, but she challenges the argument on general grounds, not in terms of the dispositionalist framework. And while Fara (2008) makes much of masking in order to address Frankfurt-style cases, his account of masks is not sensitive to either of the two principles discussed above.[23] The argument that I will advance in what follows thus offers a solution for the dispositional account of free will independent of previous dispositionalist projects.

Here, then, is Pereboom's (2014) formulation of his first case:

> A team of neuroscientists has the ability to manipulate Plum's neural states at any time by
>
> radio-like technology. In this particular case, they do so by pressing a button just before he begins to reason about his situation, which they know will produce in him a neural

[22]Both terms are McKenna's. See McKenna (2008, 2014) and Jeppsson (2020) for hard-line replies; see Demetriou (2010) and Deery & Nahmias (2017, 2022) for soft-line replies.

[23]In particular, Fara holds that an ability is masked whenever an agent tries to exercise an ability and fails to do so while their ability remains intact (Fara, 2008, 848). This account cannot allow for masks to come in degrees, as one might succeed at some action to a lesser degree than they would had their ability not been masked, and this account cannot allow for abilities to be masked while their manifestations appear.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

> state that realizes a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White. Plum would not have killed White had the neuroscientists not intervened, since his reasoning would then not have been sufficiently egoistic to produce this decision (Pereboom, 2014, 76–77).[24]

Pereboom argues that Plum1's first-order desires conform to his second-order desires and that he deliberates using his rational abilities. If Plum1's reasoning process had been different, he would have refrained from killing White. Further, Plum1 is often but not always egoistic, so his decision is ultimately in line with his character. Because of the intervention, Pereboom holds that it would seem inappropriate to hold Plum1 morally responsible for killing White. Because the neuroscientists cause the realization of just one mental state, Plum1 still exercises his rational abilities and it is in virtue of a decision that he makes that he kills White. These claims together form the grounds for an argument that none of Plum1's abilities are masked.

I nonetheless hold that this is a case of masking. The manifestation of Plum1's ability to choose on the basis of reasons appears in this case—Plum1 indeed chooses on the basis of reasons—but the degree to which that is the causal product of his own abilities is lessened by the neuroscientists' intervention. The neuroscientists, in altering Plum1's neural states such that they realize a particular reasoning process, interfere with the ordinary causal chain that otherwise would result from Plum1's attempting to deliberate about his situation.[25] In doing so, they lessen the degree to which the apparent manifestation of his ability—his making a choice on the basis of reasons—is in fact a manifestation of this ability. This does not mean that Plum1 does not genuinely make a decision, or that he is in no way the causal source of the killing. It merely means that there is a difference between the degree to which he is the causal source here and the degree to which ordinary agents are the causal sources of their own actions.[26] And this difference, given the anti-masking condition, is relevant to Plum1's satisfaction of compatibilist conditions on free will.

To this extent, Plum1's reasoning process is analogous to the cases of simultaneous masks and mimics discussed above. Recall that a disposition or an ability might be both masked and mimicked at the same time: an agent who is able to swim down a river might be caught in a fast current that carries them down the exact same path that they would otherwise have swam (we can further suppose that their arms and legs move in precisely the way that they would have if the agent were swimming of their own volition). Though the agent remains able to swim down the river, the current masks their ability and simultaneously mimics it, producing the outcome that they otherwise would have produced had they not been caught in the current.

My claim, then, is that the neuroscientists' intervention acts as both a mask and a mimic: it interferes with the ordinary causal chain in such a way that Plum1's ability to choose plays a reduced causal role in the decision that he ultimately makes and it simultaneously brings about a state that could, under different circumstances, have been the manifestation of Plum1's ability to choose on the basis of reasons. Given the anti-masking condition argued for above, this means that Plum1 does not satisfy *all* compatibilist conditions on free will and moral responsibility; there is thus a relevant difference between him and Plum4. The four-case manipulation argument thus fails.

[24]This version of Pereboom's manipulation argument is importantly different from his earlier (e.g., Pereboom, 2001) formulation insofar as in this case, the neuroscientists' manipulation occurs entirely before Plum1 begins to deliberate; in his (Pereboom, 2001) first case, Pereboom presents an agent who is manipulated on a moment-to-moment basis. For the sake of keeping this paper to a manageable length, I will only consider Pereboom's more recent version of the cases because they have been updated to deal with objections to the earlier versions of the cases.

[25]By 'ordinary causal chain' I mean the sequence that originates from Plum1 responding to and deliberating on the basis of reasons and that culminates in his making a decision and executing his intention. Cases such as Pereboom's first case in which this sequence unfolds for different reasons but nonetheless results in the same outcome are excluded from this class.

[26]Note that this point does not require any particular account of causation or another; it merely requires an account that allows causation to come in degrees.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

One might object that because Plum1 is causally determined, it is not the neuroscientists' intervention alone that masks Plum1's ability; it is the conjunction of the neuroscientists' manipulation *and* the laws' being deterministic. Regardless of what one thinks masks are, allowing for determinism to count as a mask will result in the absurd view that every time a disposition is stimulated and fails to manifest, that disposition is masked.[27] Such a position would not allow for the existence of non-surefire dispositions. If it is possible, say, for a glass to be disposed to break when dropped but luckily happen not to break on a given drop, then this position must immediately be ruled out. For this position would entail that the glass' fragility is not less-than-surefire but is instead masked by the laws' being deterministic in conjunction with the background conditions under which the glass was dropped. This is clearly not what we have in mind when we talk about fragility in everyday life, so incompatibilists who endorse this position must offer a highly counterintuitive account of modal properties.[28]

Further, one might object that Plum1 *loses* his ability in this case rather than having it masked: perhaps the neuroscientists' intervention simply robs him of the ability to make a choice here. But this would be highly counterintuitive for two reasons. First, it seems intuitively clear that there is at least some degree to which Plum1's own agential abilities play a role in his choosing to kill White, a position which could not be explained if Plum1 were entirely robbed of his abilities. Second, this does not cohere with our intuitions about other cases involving simultaneous masks and mimics. Recall the case in which the Hater of Styrofoam stomps on a glass protected in styrofoam just as the glass makes contact with the ground. If we hold that Plum1's ability is removed rather than masked, then there is no principled reason to hold that the glass' fragility is not removed as well. But this is clearly wrong: the glass remains fragile even though its fragility does not cause it to break.

Pereboom's cases are constructed such that we intuitively accept that there is no difference between Plum1 and Plum2, Plum2 and Plum3, and Plum3 and Plum4; this sequence supports the *No Difference* premise. Advocates of the argument might thus reasonably ask where the analogy between one Plum and the next fails. As the second case is typically drawn up, Plum2 is just like any other agent except that he was programmed at birth such that he would kill White in this particular case through his own decision-making procedures. The view that I have advanced here thus predicts that Plum2 is responsible for his actions; given the assumption that he was designed in such a way that he is just like any other agent, his abilities are not masked in the way that Plum1's are: unlike Plum1, his decisions and actions are entirely the manifestations of the abilities relevant to his free will and responsibility.[29]

I think that this is the correct conclusion to draw about this case. The power of the four-case argument stems from the fact that the intuition that Plum1 is not responsible for killing White is very strong. Left only with the intuition that Plum2 is not responsible for his actions, the argument is much weaker; dispositional compatibilists can draw on extant replies to, for instance, Mele's (2006) zygote argument as a means of eliciting their preferred intuitions about this case. While advocates of the four-case argument might attempt to hold onto the claim that the temporal difference between the manipulation in the first and second cases is irrelevant to the facts about responsibility, the anti-

[27]Further, in the case of dispositions such as loquaciousness that do not have specific stimulus conditions, this position yields the result that dispositions are *always* masked. And it is absurd to hold that someone's loquaciousness is masked by the past and the laws' being deterministic *whenever* someone does not talk.

[28]Incompatibilists might, of course, offer independent arguments against this position and hold that it really is the case that there are no non-surefire dispositions. However they argue for this position, they will be forced to accept highly counterintuitive positions. Given that most arguments for incompatibilism depend heavily on their tracking our ordinary intuitions, this would be a major theoretical strike against any such position.

[29]This depends, of course, on precisely how the details of the second case are spelled out. Advocates of the four-case argument might construct a second case in which the programming does somehow act as a mask, in which case the line would be drawn between the second and third cases. And similarly, if the third case is detailed in such a way that Plum3's egoistic background somehow masks his abilities, then the line can be drawn between the third and fourth cases.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

masking condition that I have advanced here provides us with just the resources to draw this difference. When manipulation of the kind described in the first case occurs, a mask mitigates responsibility; when programming of the kind described in the second case occurs, no mask is present and there is a metaphysical difference between the two cases that can explain why one Plum is blameworthy and the other is not.[30]

## 5. Causal Histories and Agency-Preserving Interferences

Advocates of the four-case argument can push back against the reply that I have developed here in two ways. First, they can attempt to revise the first case to accommodate my proposed compatibilist condition on free will, denying any metaphysical difference between Plum1 and Plum4. Second, they can further push the argument that masks do not undermine responsibility by presenting cases in which agents' abilities are masked but agents nonetheless intuitively seem responsible for their actions.[31] Let us consider these arguments in turn.

First, consider how one might alter the first case in such a way that the relevant metaphysical difference no longer exists. The most natural way, of course, is to develop a version of the case in which Plum1's abilities are not masked but in which the intuition that Plum1 is not responsible for his actions remains in place. One way of doing so is to alter the case such that the neuroscientists *alter* Plum1's abilities rather than interfering with their manifestations. Assume, following the argument that I have developed here, that Plum4's abilities are not masked and call the neural state in virtue of which Plum4 is able to choose and act freely *N*. Can the neuroscientists implant *N* into Plum1, thus altering his dispositions rather than masking them?[32] If so, then there is a version of the case in which Plum1's dispositions are not masked but are instead altered.

But this line of argument does not succeed. Suppose first that in implanting *N* in Plum1, they alter his dispositions. If so, the manipulation featured in the case is far more intrusive than the form of manipulation that Pereboom describes in his (Pereboom, 2014) version of the first case. It is important to note here that the reason that Pereboom describes such a subtle form of manipulation —the mere enhancement of an egoistic reasoning disposition—is because earlier versions of his argument (e.g., his [Pereboom, 2001] first case) were themselves targets of significant compatibilist critiques. In his (Pereboom, 2001) version of the cases, the neuroscientists controlled Plum1's neural states on a moment-to-moment basis, realizing one neural state after the other in such a way that he killed White. Fischer (2004), Mele (2005), and Demetriou (2010) each raised objections to this version of the cases on the grounds that such a drastic form of manipulation undermines agency altogether. Implanting a different set of dispositions in an agent would encounter many of the same difficulties. Pereboom describes a subtler form of manipulation in his (Pereboom, 2014) in order to avoid these worries; if the only way of avoiding the objection that I have raised here is by returning to a much stronger form of manipulation, then the argument will fail to satisfy other compatibilist criteria on free will and fail for different reasons.

Advocates of the four-case argument might object that there must be some minimal degree of manipulation such that Plum1's abilities are not masked. Perhaps, for instance, making a decision on the basis of one's own abilities is an *intrinsic current time-slice notion*, determined entirely by Plum1's neural states at a time. If so, then there must be some way of manipulating Plum1 such that he still makes a decision entirely on the basis of his abilities.

While this seems like a promising line of argument, I think that it is ultimately mistaken. Making a decision on the basis of one's own abilities is not determined entirely by one's neural states at a

[30]This position is not without precedent. Fara (2008) builds a temporal element into his account of intrinsic masks as a means of distinguishing between Frankfurt's (1969) Jones and Frankfurt's (1971) willing addict.

[31]I am not hereby equating responsibility and agency. What matters for my position is that abilities being masked sometimes results in diminished responsibility; this latter objection denies this claim.

[32]Thanks to an anonymous referee for raising this worry.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

time, but is instead partly determined by one's history. To see why this should be the case on the dispositional account of free will, consider how the thesis that deciding on the basis of one's abilities is an intrinsic current time-slice notion would generalize to other dispositions—if this is thesis is true, after all, we should expect it to be true of all similar modal properties, not just of abilities. With respect to fragility, for instance, this thesis predicts that an object's breaking due to its fragility is an intrinsic current time-slice notion. But this is implausible in light of the above-discussed case of the simultaneous mask and mimic. Suppose that two glasses are perfect duplicates of each other. Glass A is protected with a piece of styrofoam and glass B is not. Both are in the presence of the Hater of Styrofoam and both are dropped simultaneously. Both glasses break—A because the Hater of Styrofoam smashes it as soon as it strikes the ground and B because it is fragile and has been dropped.[33] Even if (by some miracle) the two glasses break in precisely the same way, such that they remain perfect intrinsic duplicates as they break, A's breaking is the product of a mask and a mimic while B's breaking is the manifestation of its fragility. This kind of case undercuts the suggestion that a disposition's manifesting is entirely determined by its intrinsic states at a time.

Generalizing this conclusion back to the case of agents' abilities helps demonstrate that whether an ability manifests is not determined entirely by an agents' intrinsic states at a time. It is possible that there is some perfect intrinsic duplicate of Plum1 whose abilities manifest in the ordinary sequence and who is entirely responsible for his actions, but the fact that there is some such possible duplicate does not entail that Plum1 himself decides on the basis of unmasked abilities. In virtue of this, the position that I have developed here is a version of historical compatibilism,[34] albeit one based on general principles about the nature of modal properties and their manifestations rather than our intuitions about responsibility alone.

A second way in which advocates of the four-case argument can push back against the argument that I have developed here is by rejecting the anti-masking condition on responsibility. Pereboom, in responding to soft-line compatibilists who hold that Plum1 lacks agency altogether, argues that "[a]gency is regularly preserved in the face of certain involuntary momentary external influences" (Pereboom, 2014, 76). He offers the example of a sports fan who finds out that the home team lost the big game. If the fan's worst character traits were amplified by their disappointment in the loss, we would still take them to exercise their agency if they did something reprehensible. And while we might ordinarily take such a person to be morally responsible for their conduct, Pereboom argues, the first manipulation case presents a situation in which agency is preserved, but responsibility is not.

If it is possible that someone who is subject to influences that appear to undermine their rational abilities to nonetheless remain intuitively fully in control of their actions in a way that would make them responsible, as seems to be the case here, then there are grounds to reject the anti-masking condition on blameworthiness that I proposed above. But there are two replies available on behalf of the dispositional view.

First, dispositionalists can appeal to considerations about tracing to distinguish such cases from cases in which masks undermine responsibility: perhaps, despite the fact that the sports fan's abilities are masked at the time at which he makes his immoral decision, he is responsible for his conduct in virtue of earlier decisions that he made. Such an approach would supplement the tracing condition introduced in Section 3.

While there is a surface-level similarity between Plum1 and the sports fan insofar as some of each of their abilities are masked, there is thus a relevant difference between them: the sports fan is responsible for the presence of the mask while Plum1 is not. The sports fan could have taken steps to

[33]One might worry that there is still some sense in which glass A breaks because it is fragile. But we can instead consider finer-grained dispositions such as the disposition to break when dropped specifically; this disposition clearly does not manifest when a glass breaks because it is stepped on.

[34]See Fischer and Ravizza (1998) and Mele (2019) for other historical accounts of blameworthiness.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

prevent themself from falling under the influence of their fandom and acting immorally and should have foreseen the possibility that they might do so. On this line, the difference between Plum1 and the sports fan is analogous to the difference between a drunk driver who voluntarily got drunk without taking any precautions to keep themself from driving and a drunk driver who was force-fed alcohol and involuntarily placed behind the wheel. While both drunk drivers have their abilities masked, only the driver whose abilities' being masked can be traced to earlier decisions of theirs can properly be held responsible.

One might object that the same is true of Plum1, as he could have taken steps to prevent himself from being manipulated into killing White. But the difference here, of course, is that Plum1 could not reasonably be expected to have known that he would be manipulated into killing White, whereas it is reasonable to expect a sports fan to know that they might become upset if their team loses or to expect anyone to know that they might drive if they have too much to drink without taking steps to stop themself from driving. And while one might hold that Plum1 is similar to either such person insofar as he put himself into a position in which he could easily be manipulated into killing White, adopting such a line would undermine the intuitive pull of the manipulation argument and would, in itself, threaten the *No Responsibility* premise.

A second theoretical option here is to appeal to degrees of masking and degrees of responsibility. We can accept that the sports fan's abilities were masked and that he is intuitively blameworthy for his conduct, but hold that the presence of this mask reduces the degree to which he is blameworthy for his actions. As discussed above, masks are gradable; a thick sheet of bubble wrap does more to prevent a fragile glass from breaking than a thin sheet does, despite the fact that both are masks. Plausibly, the sports fan's ability is merely masked to some degree, such that he remains responsible for his conduct despite the fact that he in fact manifests his abilities to choose. If we take this line, we can grant Pereboom that the sports fan's agency is preserved and that he is the subject of a momentary external influence, but hold that his agency is preserved only to a lesser degree than a normal agent's is.

## 6. The Dialectic

Let us turn now to the dialectical role of the argument that I have advanced here. There are many compatibilist objections to the four-case argument. Precisely what does the argument that I have offered here add? I think that my view presents several advantages. First, the most obvious means of undermining any compatibilist reply to the four-case argument is to alter the details of the case such that whatever proposed compatibilist condition is purported not to be met in the original version of the case is now met. But the argument that I have advanced here, as noted above, cannot easily be circumvented by this maneuver. If my argument succeeds, the core structure of the four-case argument involves a violation of a proposed compatibilist condition on free will: as long as Plum1 is actively manipulated by neuroscientists, his decision-making abilities are masked. There is thus no easy way for advocates of the four-case argument to hold that Plum1 satisfies all compatibilist conditions on responsible agency while eliciting the intuition that he is not responsible for his conduct.

While my aim in this paper has been to show that dispositional compatibilists have the resources to counter the four-case argument, compatibilists who reject the dispositional view might still be able to adapt the argument that I have developed here. If abilities can be masked in the same way that dispositions can, then regardless of whether the new dispositionalism is the right view of free will, the arguments that I have advanced here can be taken on board by other compatibilists. Any compatibilist who grants that abilities can be masked can hold that there is a difference between the first two cases that undermines Pereboom's *No Difference* premise.

Further, note that the view that I have defended here, like other components of the dispositional account, begins from clear intuitions about simple modal properties and then moves to more complicated considerations about agents' abilities. The view that I have defended here is independently

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

motivated by considerations about dispositions and thus presents a case for rejecting the manipulation argument that does not depend primarily on our intuitions about free will and responsibility in particular.

One might worry that the view that I have defended here focuses too much on Pereboom's *No Difference* premise. The argument that I have developed is a soft-line reply insofar as it centers around rejecting the *No Difference* premise. But any soft-line compatibilist must take the hard-line reply with respect to certain cases. As discussed in Section 4, I think that the *No Difference* premise is true when applied to the second case: there is no difference relevant to responsibility between an agent programmed at birth in the way that Pereboom describes and an ordinary causally determined agent like Plum4. The view that I have defended here, however, also involves a component of a hard-line response with respect to the first case insofar as I hold that Plum1 is (or at least in principle can remain) partially blameworthy for killing White insofar as there is a degree to which his actions are partially the manifestations of his own rational abilities.

Some manipulation arguments (e.g., Mele's [2006] and Todd's [2011]) begin from cases involving an agent manipulated at birth, rather than an actively manipulated agent.[35] One might thus wonder what dialectical role the argument I have advanced here plays, as I have taken a hard-line stance towards cases that others use to motivate manipulation arguments. As noted above, part of what makes the four-case argument so powerful is that the intuition that Plum1 is not responsible is so strong. Indeed, I find this intuition more powerful than I find the intuition that Plum2 is not responsible; it is only because I have the relevant intuition about the first case that I find that manipulation arguments work. I thus think that an argument that undermines the bridge from the first case to the second does much of the necessary work relevant to cut off manipulation arguments. Even if one's intuitions about these cases differ from mine, however, note that manipulation arguments are a varied class. Showing why one manipulation argument fails might not show why all other manipulation arguments fail, but if different manipulation arguments fail for different reasons, then a pluralistic set of responses is necessary to undermine the entire class of arguments.

## 7. Conclusion

The four-case manipulation argument has long been one of the most influential challenges to compatibilism. I have offered a novel reply to the argument here by arguing that Plum1's rational abilities are masked in the first case, making for a metaphysical difference between Plum1 and Plum4. Though components of what I have argued here cohere with other compatibilist projects, the argument that I have advanced fits best with dispositional compatibilism. Beyond securing our freedom from the challenge presented by van Inwagen's Consequence Argument in a way that allows compatibilists to say that we remain able to do otherwise in a deterministic universe, the new dispositionalism affords compatibilists a principled means of undercutting the threat that Pereboom's manipulation argument has long posed.

**Acknowledgments.** Thanks to Eddy Nahmias, Shaun Nichols, Joseph Orttung, Derk Pereboom, Geoffrey Weiss, and two anonymous referees at this journal for many helpful comments on this paper.

**Gus Turyn** is a Ph.D. candidate in philosophy at Cornell University. His research focuses on moral psychology, metaphysics, and epistemology.

## References

Ayer, A.J. (1954). Freedom and necessity. In *Exploring philosophy: An introductory anthology*. Palgrave Macmillan.

Bird, A. (1998). Dispositions and antidotes. *The Philosophical Quarterly*, 48(191), 227–234.

[35]See e.g., Tierney (2013), Khoury (2014), and Deery and Nahmias (2017, 2022) for some replies to arguments that begin with cases more similar to Pereboom's second case.

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press

Choi, S. (2011). What is a dispositional masker? *Mind*, 120(480), 1159–1171.
Choi, S. (2017). Against intrinsic interferers: A critique of kittle. *The Philosophical Quarterly*, 67(269), 845–854.
Clarke, R. (2009). Dispositions, abilities to act, and free will: The new dispositionalism. *Mind*, 118(470), 323–351.
Coates, D. J., & Swenson, P. (2013). Reasons-responsiveness and degrees of responsibility. *Philosophical Studies*, 165(2), 629–645.
Cyr, T. W. (2016). The parallel manipulation argument. *Ethics*, 126(4), 1075–1089.
Deery, O., & Nahmias, E. (2017). Defeating manipulation arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies*, 174(5), 1255–1276.
Deery, O., & Nahmias, E. (2022). Why the manipulation argument fails: Determinism does not entail perfect prediction. *Philosophical Studies*, 180(2), 451–471.
Demetriou, K. (2010). The soft-line solution to Pereboom's four-case argument. *Australasian Journal of Philosophy*, 88(4), 595–617.
Fara, M. (2008). Masked abilities and compatibilism. *Mind*, 117(468), 843–865.
Fischer, J. M. (2004). Responsibility and manipulation. *The Journal of Ethics*, 8(2), 145–177.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control*. Cambridge University Press.
Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23), 829–839.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68(1), 5–20.
Gebharter, A., & Fischer, F. (2021). A causal Bayes net analysis of dispositions. *Synthese*, 198(5), 4873–4895.
Janzen, G. (2016). 'Brain-malfunction' cases and the Dispositionalist reply to Frankfurt's attack on PAP. *Australasian Journal of Philosophy*, 94(4), 646–657.
Jeppsson, S. (2020). The agential perspective: A hard-line reply to the four-case manipulation argument. *Philosophical Studies*, 177(7), 1935–1951.
Johnston, M. (1992). How to speak of the colors. *Philosophical Studies*, 68(3), 221–263.
Khoury, A. (2014). Manipulation and mitigation. *Philosophical Studies*, 168(1), 283–294.
Lewis, D. (1997). Finkish dispositions. *The Philosophical Quarterly*, 47(187), 143–158.
Manley, D., & Wasserman, R. (2007). A gradable approach to dispositions. *The Philosophical Quarterly*, 57(226), 68–75.
Manley, D., & Wasserman, R. (2008). On linking dispositions and conditionals. *Mind*, 117(465), 59–84.
Martin, C. B. (1994). Dispositions and conditionals. *The Philosophical Quarterly*, 44(174), 1–8.
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. (2014). Resisting the manipulation argument: A hard-liner takes it on the Chin. *Philosophy and Phenomenological Research*, 89(2), 467–484.
Mele, A. R. (2005). A critique of Pereboom's 'four-case argument' for incompatibilism. *Analysis*, 65(1), 75–80.
Mele, A. R. (2006). *Free will and luck*. Oxford University Press.
Mele, A. R. (2019). *Manipulated agents*. Oxford University Press.
Pereboom, D. (1995). Determinism al dente. *Noûs*, 29(1), 21–45.
Pereboom, D. (2001). *Living without free will*. Cambridge University Press.
Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.
Pereboom, D. (2020). *Wrongdoing and the moral emotions*. Oxford University Press.
Pereboom, D. (2022). *Free will*. Cambridge University Press.
Sartorio, C. (2016). *Causation and free will*. Oxford University Press.
Sartorio, C. (2023). *Causalism*. Oxford University Press.
Shabo, S. (2010). Uncompromising source incompatibilism. *Philosophy and Phenomenological Research*, 80(2), 349–383.
Smith, M. (2003). Rational capacities, or: How to distinguish recklessness, weakness, and compulsion. In S. Stroud, & C. Tappolet (Eds.), *Weakness of will and practical irrationality*. Clarendon Press.
Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 187–211.
Tierney, H. (2013). A maneuver around the modified manipulation argument. *Philosophical Studies*, 165(3), 753–763.
Todd, P. (2011). A new approach to manipulation arguments. *Philosophical Studies*, 152(1), 127–133.
Todd, P. (2016). Manipulation arguments and the freedom to do otherwise. *Philosophy and Phenomenological Research*, 95(2), 395–407.
Turyn, G. (2021). On dispositional masks. *Synthese*, 199(5–6), 11865–11886.
van Inwagen, P. (1983). *An essay on free will*. Oxford University Press.
Vetter, B. (2019). Are abilities dispositions? *Synthese*, 196(1), 201–220.
Vihvelin, K. (2004). Free will demystified: A dispositional account. *Philosophical Topics*, 32(1/2), 427–450.
Vihvelin, K. (2013). *Causes, laws, and free will: Why determinism doesn't matter*. Oxford University Press.
Whittle, A. (2010). Dispositional abilities. *Philosophers' Imprint*, 10(12), 1–23.

**Cite this article:** Turyn, G. 2025. Masked Abilities and the Four-Case Manipulation Argument. *Canadian Journal of Philosophy* 55: 84–98, doi:10.1017/can.2025.10044

https://doi.org/10.1017/can.2025.10044 Published online by Cambridge University Press