Spring 5-14-2021

Masked Abilities and Manipulation Arguments

Matthew Turyn

MASKED ABILITIES AND MANIPULATION ARGUMENTS

by

MATTHEW TURYN

Under the Direction of Eddy Nahmias, PhD

ABSTRACT

Dispositional accounts of free will provide a promising avenue for compatibilists about free will and determinism to respond to manipulation arguments, but requires an adequate account of masked and finkish abilities. In this thesis, I argue that an account of masked dispositions that addresses the context sensitivity, gradability, and dispositional quality of masks can allow dispositional compatibilists to level viable responses to various manipulation arguments. I address my response towards both Pereboom's and Mele's manipulation arguments.

INDEX WORDS: Free will, Dispositions, Manipulation arguments

MASKED ABILITIES AND MANIPULATION ARGUMENTS

by

MATTHEW TURYN

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2021

MASKED ABILITIES AND MANIPULATION ARGUMENTS

by

MATTHEW TURYN

Committee Chair: Eddy Nahmias

Committee: Edward Cox

Andrea Scarantino

Neil Van Leeuwen

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2021

1. ACKNOWLEDGEMENTS

Thanks to all of the professors who helped me develop the ideas that I present in this paper. I am especially grateful to Eddy Nahmias, whose willingness to look over early drafts and meet countless times to discuss these ideas helped me develop my own view and learn how to present it. I am also incredibly grateful to my friends and family who looked over earlier drafts and helped me through this process, especially Ryan Belle and my mother, Anne Turyn.

TABLE OF CONTENTS

1 INTRODUCTION

Dispositional accounts of free will such as Kadri Vihvelin's (2004; 2013) offer a promising method for compatibilists to respond to manipulation arguments such as Derk Pereboom's (2001; 2014) and require further attention than they have so far garnered. Pereboom has argued that *causal determinism*, the thesis that all future events are necessitated by prior events and the laws of nature, precludes the possibility of free will and moral responsibility by means of a series of thought experiments. In this thesis, I develop Vihvelin's response to Pereboom and defend it against possible objections.

I will argue that free will and moral responsibility are *compatible* with determinism. Other *compatibilists* have argued that particular conditions are necessary for free will. 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 proposes a series of analogies between decreasingly controlled agents who purportedly do not differ with respect to moral responsibility. Because each agent in Pereboom's thought experiment lives in a deterministic universe, their actions are causally determined. I argue that Pereboom's argument fails, and that free will and moral responsibility are compatible with causal determinism. Each case is set up such that Plum is supposed to satisfy any and all proposed compatibilist conditions for free will. Each Plum has some egoistic reason to want

White dead, and each ultimately kills White. In the first case, Plum1 is actively manipulated by neuroscientists who change his reasoning process such that he chooses to kill White. In the second case, Plum2 is designed by neuroscientists at birth such that it is causally determined that he will kill White when the opportunity arises. Plum3 is raised in an environment in which his egoistic tendencies are stoked and develops the characteristics that lead to him being causally determined to kill White when the opportunity arises. Plum4 is a normal agent in a deterministic universe who reasons egoistically when presented with the opportunity to kill White. Pereboom argues that nothing relevant to moral responsibility distinguishes one case from the next and that the best explanation for their common lack of responsibility is that all are causally determined by forces over which they ultimately lack control. Pereboom's argument thus rests on two premises: (1) there is no difference relevant to moral responsibility between Plum1, Plum2, Plum3, and Plum4, and (2) Plum1 is not morally responsible. If the best explanation for why each Plum is not responsible is because each is causally determined, then no causally determined agent can ever be morally responsible. Compatibilists about determinism and moral responsibility thus must reject one of these premises. Those who reject the first premise offer a 'soft-line' response; those who reject the second offer a 'hard-line' response.[1]

To defend the compatibility of free will and determinism, I will focus on the conditions for *masked abilities*. According to Vihvelin's dispositional account of free will, agential abilities such as the ability to speak or solve puzzles are complex dispositions. Just as a fragile glass is disposed to break when it is dropped, an agent is disposed to make decisions. The *stimulus condition* for a glass' disposition to break is the glass' being dropped; the stimulus condition for

[1] Both terms are McKenna's. See McKenna (2006, 2014) for a hard-line response; see Demetriou (2010), Fischer (2014), and Deery & Nahmias (2017) for soft-line responses.

an agent's disposition to make decisions is the agent's trying to make a decision. When the glass is dropped and breaks, its disposition *manifests*; when an agent tries to make a decision and succeeds at doing so, their ability is *exercised*.

Much in the same way that a glass can still be fragile even if it is wrapped in bubble wrap (Johnston 1992), an agent can still be disposed to make decisions even if she is temporarily under the influence of something that will prevent her from making decisions. Entities such as bubble wrap and malicious neuroscientists with mind control devices are *masks* (some philosophers, e.g. Bird 1998, use the term 'antidotes').

This thesis is structured as follows. In section 2, I discuss the roots of dispositional accounts of free will and recent accounts of dispositions. In section 3, I discuss masked abilities and Michael Fara's (2005) analysis of masks. I argue that his analysis is unsuccessful and propose a new analysis. In section 4, I use my account of masks to develop a dispositional response to Pereboom's (2001) manipulation argument. I argue that Pereboom's argument relies on comparing agents whose rational abilities are masked to agents whose rational abilities are not masked. Plum1's rational abilities are masked, whether Plum2's and Plum3's rational abilities are masked depends on how we read the details of each case, and Plum4's rational abilities are not masked. Pereboom's first premise, that there is no difference between each Plum, is false. In section 5, I take on an objection that Pereboom could level against my dispositional response to manipulation arguments and argue that his more recent manipulation cases (2014) do not avoid the problems with his (2001) cases. In section 6, I argue that my response also addresses Alfred Mele's zygote argument, a modified manipulation argument which Mele argues is able to circumvent the problems with Pereboom's argument.

2 COUNTERFACTUAL ANALYSES AND THE DISPOSITIONAL ACCOUNT OF FREE WILL

Kadri Vihvelin's dispositional account of free will has its roots in classical compatibilist accounts. For Vihvelin, agential abilities are "bundles of dispositions" (2004; 2013). Fragility is a simple disposition, but the complex dispositions that agents possess, such as the abilities to speak and to choose, are combinations of simpler dispositions. An account of agential abilities thus requires an account of dispositions; Vihvelin endorses a conditional analysis according to which possession of a disposition depends on the truth value of a counterfactual conditional. Something is fragile just in case it would break if it were struck; an agent is able to act in some way just in case her trying to act in that way would cause her to do so. A glass has a disposition (fragility) that manifests when it is exposed to its stimulus conditions (being struck or dropped); an agent has an ability (the ability to choose on the basis of reasons) that manifests when the stimulus conditions come about (the agent trying to choose). Vihvelin's account of dispositions is well-equipped to address two key problems with other accounts of dispositions: that an entity need not exercise its disposition to the same degree each time it is exposed to the relevant stimulus conditions and that dispositions are impermanent.

Classical compatibilists (Hume 1748 and Moore 1912, for instance) argued that an agent has free will just in case she would have done otherwise had she decided to do otherwise. Determinism thus does not rule out free will; in many cases, causally determined agents would have done otherwise had they chosen to do so, which classical compatibilists argue agents are able to do. Such an account requires what Lewis calls the 'simple conditional analysis' (SCA) of a disposition:

> 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).

According to the SCA, a glass is disposed at time *t* to break upon being struck or dropped iff, if the glass were to be struck or dropped at time *t*, it would break. Similarly, an agent is disposed at some time to act in some way iff, if the agent were to choose at that time to act in that way, she would act in that way. But this analysis fails to account for 'finkish' dispositions. First discussed by CB Martin (1994), a fink is some phenomenon that would alter an object's disposition if that disposition's stimulus conditions came about. Lewis's (1997) example describes a wizard who is especially fond of a particular glass. Though the glass is fragile—disposed to break when struck—the wizard keeps watch over it and, if anything were to strike the glass, he would cast a spell that would cause the glass to cease to be fragile. That the glass ceases to be fragile when the wizard casts the spell does not mean that the glass was not fragile at earlier moments, insofar as it possessed a particular intrinsic property in virtue of its molecular structure. The glass' disposition to break when struck is thus finkish, but until the stimulus conditions come about, the glass remains fragile. As Vihvelin puts it, "to be at *risk* of losing our abilities is not the same thing as *actually* losing our abilities" (2004: 448; Vihvelin's emphasis).

Lewis developed an analysis of dispositions intended to circumvent the problem posed by finks by specifying that an entity must retain the intrinsic property (some fact about the internal structure of the entity) in virtue of which it has a disposition in order to act in the manner that it is so disposed:

> Something *x* is disposed at time *t* to give response *r* to stimulus *s* iff, for some intrinsic property *B* that *x* has at *t*, for some time *t′* after *t*, if *x* were to undergo stimulus *s* at time *t* and retain property *B* until *t′*, *s* and *x*'s having of *B* would jointly be an *x*-complete cause of *x*'s giving response *r* (Lewis 1997: 157).

The wizard's favorite glass is disposed at time *t* to break upon being dropped iff, for the molecular structure of the glass at time *t*, for some time *t′* after *t*, if the glass were dropped at *t*

and retained its molecular structure until *t′*, the glass' being dropped and the glass' molecular structure would together cause the glass to break. Because the wizard intervenes, the glass does not retain the molecular structure in virtue of which it had previously been fragile. But if the wizard had not intervened, the glass would have broken when struck. Lewis's analysis would correctly yield the conclusion that the glass is fragile at the time before the spell is cast.

While Lewis's analysis does well to solve the problem presented by finks, he was unable to account for a similar issue: the problems presented by masks. David Manley and Ryan Wasserman (2008) offer a simplified account of dispositions, which they argue avoids that problem. Masks are entities or forces that bear some relation to an object such that the object would not exercise its disposition if it were to be exposed to its stimulus conditions despite retaining the intrinsic property in virtue of which the object is so disposed. A fragile glass, for instance, can have its fragility masked by being placed in a layer of protective bubble wrap (Johnston 1992). While the glass is still fragile—its intrinsic properties are such that under normal conditions, it would break when struck—its present circumstances are not conducive to the glass' disposition being exercised.

Such cases seem to provide counterexamples to Lewis's conditional analysis of dispositions. Suppose that the glass in question were dropped while wrapped in bubble wrap. For the time period in question, it would retain the intrinsic property in virtue of which it is otherwise disposed to break when dropped—in this case, its molecular structure—but its having that property and being exposed to the stimulus that would otherwise cause it to break would, in this case, not cause it to break.

Lewis attempts to skirt this objection through means of what Manley and Wasserman call "the strategy of getting specific" (2008: 66). On this strategy, we should take it that the glass is

not simply disposed-to-break-when-dropped, but is instead disposed-to-break-when-dropped-while-not-wrapped-in-bubble-wrap. For Lewis, solving the problem that masks present is simply a matter of making sure that we correctly identify the specific disposition in question and the specific stimulus conditions for that disposition to manifest.

Manley and Wasserman argue against this strategy. There are infinitely many possible masks for any given disposition, so no finite specifications of conditions can ensure that there are no masks present (2008: 64). They also note that we cannot simply appeal to ideal conditions, as Mumford (1998) suggests we do, because this would result in a tautology: all objects are disposed to break under conditions under which they would break (Manley and Wasserman 2008: 65). Manley and Wasserman advocate an understanding of dispositions according to which dispositions are 'gradable,' meaning that one object can have more or less of a particular disposition than another. A glass is disposed to break when dropped even if it does not break on any given drop; a glass that breaks on only 80% of drops would still be considered fragile. Some glasses are more fragile than others, but a glass that breaks on 95% of drops and a glass that breaks on 60% of drops are both fragile. Two glasses that differ in fragility differ in degree of their common disposition, not in kind, a fact that Lewis's strategy does not capture. In light of this problem, Manley and Wasserman propose the following analysis, which they call PROP:

> *N* is disposed to *M* in *C* if and only if *N* would *M* in some suitable proportion of *C*-cases (Manley and Wasserman 2008: 76).

Their analysis is plausible and offers a degree of simplicity that Lewis's does not, but it sacrifices the intricacies of Lewis's account.

2.1 Vihvelin's account of dispositions

Vihvelin builds on the work of earlier dispositional analyses, focusing on the ability to make choices. She offers the following analysis, which she calls LCA-PROP-ABILITY, a nod to Lewis's (1997) account and to Manley and Wasserman's (2008) account:

> *S* has the narrow ability at time *t* to do *R* as the result of trying iff, for *some* intrinsic property *B* that *S* has at *t*, and for some time *t′* after *t*, if *S* had the opportunity at *t* to do *R and S* tried to do *R* while retaining property *B* until time *t′*, then in a *suitable proportion of these cases*, *S*'s trying to do *R* and *S*'s having of *B* would be an *S*-complete cause of *S*'s doing *R* (Vihvelin 2013: 187; Vihvelin's emphasis).

Consider how this would apply to a baseball player's ability to hit a home run. The best baseball players are able to hit home runs in less than a tenth of their attempts (or 'at bats'). A baseball player is thus able to hit a home run at time *t* as the result of trying iff, for the properties that he has at *t*, and for some time *t′* after *t*, if he had the opportunity to hit a home run and tried to hit a home run while retaining his intrinsic properties, then in a suitable proportion of cases, his trying to hit a home run and his having a particular set of intrinsic properties would jointly cause him to hit a home run. Even if a player failed to hit a home run in one given at bat, he might still be able to hit a home run.

Possession of the ability to choose, which is a bundle of dispositions, is not sufficient for free will. Instead, agents must also be in a position to successfully exercise the ability. Central to Vihvelin's account is the distinction between "narrow" and "wide" abilities (Vihvelin 2013: 11). Narrow abilities are those which an agent has in virtue of her intrinsic properties; wide abilities are those which an agent has in virtue of her intrinsic properties *and* her surroundings. Whether an agent has a *narrow* ability to choose on the basis of reasons depends solely on facts about her,

including her rational capacities. Whether an agent has the *wide* ability to choose depends on her narrow abilities in conjunction with her surroundings.

Vihvelin's analysis accounts for the fact that dispositions are gradable. As long as one succeeds at *A*-ing in a suitable proportion of cases, one is able to *A*. For different abilities, the threshold for possession of the ability might be higher or lower. Baseball players who are only able to hit a home run in every twentieth or thirtieth at bat are still able to hit home runs. Without accounting for gradability, we would have to conclude that for any given at bat during which a player does not hit a home run, the player was unable to do so, a concession that would undermine a dispositional analysis of abilities. Because hitting a home run is so difficult, the number of cases in which one must succeed in order to have the ability is relatively low. Note that the player's ability differs in degree between game conditions and other, friendlier conditions such as batting practice, in which they might be expected to do so more frequently. We can thus distinguish between the ability-to-hit-home-runs-in-games and the ability-to-hit-home-runs-in-batting-practice.[2]

In her response to Harry Frankfurt's (1969) attack on the 'Principle of Alternative Possibilities,' Vihvelin (2004, 2013) discusses how a dispositional analysis of free will relates to finkish dispositions.[3] Frankfurt describes Jones, an agent who is shadowed by a neuroscientist, Black, while Jones decides whether to kill White. If Jones decides not to kill White, Black will intervene by means of a device that he has planted in Jones's brain, causing him to kill White. As

[2] Whittle (2010) makes use of these more specific abilities in her argument against dispositional compatibilism. Whittle calls context-specific abilities 'local abilities' and the more broad abilities 'global abilities.' See note 10 for further discussion of Whittle's argument.

[3] Smith (2003) and Fara (2005, 2008) offer responses to Frankfurt that resemble Vihvelin's. See section II of this paper for a further discussion of Fara's account of masks. See section III for further discussion of Vihvelin's response to Frankfurt.

it happens, Jones chooses of his own accord to kill White, so Black never intervenes. Frankfurt concludes that one can be morally responsible for their actions even if one could not have done otherwise. Black, the would-be intervener, never *actually* intervenes with Jones's abilities. Because he never actually intervenes, he merely threatens to remove Jones's ability to choose (and to do) otherwise. Jones's disposition to choose is finkish in the same way that the wizard's glass is finkishly disposed to break when struck. Because the glass is never struck, the wizard never removes the glass' disposition to break when struck. Because Jones does not decide against killing White, Black never removes Jones's ability to choose (and to do) otherwise.

Vihvelin's response to Frankfurt highlights the approach that she takes towards finkish dispositions. Her response to masks is similar; when some environmental factor prevents an agent's ability from being exercised in a suitable proportion of cases, the agent lacks the wide ability. Related to masks are mimics, which are cases in which an object without a disposition acts in the way that it would if it did have that disposition. Manley and Wasserman (2008) cite the example of a brick that is not fragile except in cases in which it is dropped on one particular point on its surface. While the brick is not likely to break when dropped, it will break when dropped on that particular spot. Similarly, I am not able to speak Russian because I am not disposed to speak Russian as a result of my trying to do so. I have heard enough Russian in movies to know roughly what Russian sounds like, though, so there is a slim chance that if I were to babble some Russian-sounding syllables, I would say something intelligible in Russian. But my saying so would merely mimic a Russian speaker's genuine ability.

I will assume for the remainder of this thesis that Vihvelin's account of agential abilities is correct. By understanding abilities as gradable, impermanent, and susceptible to finks and to masks, we can see why manipulation arguments fail to depict manipulated agents whose abilities

to choose on the basis of reasons are analogous to agents who are not manipulated but are only causally determined.

2.2 Vihvelin's response to manipulation arguments

Before I propose my response to manipulation arguments such as Pereboom's (2001), it should be noted that Vihvelin, too, has argued that manipulation arguments fail to undermine compatibilist accounts of free will. On Vihvelin's account, to have free will is to have the wide ability to choose on the basis of reasons. In her response to Pereboom's manipulation argument (2013: 148-155), Vihvelin questions whether the details of the cases are such that what holds of the first case (in which the neuroscientists actively intervene) also holds of the fourth (in which the agent is not manipulated but is merely causally determined). Vihvelin argues that because Plum1's mental states are not properly caused, he cannot make rational choices. There is thus a difference relevant to moral responsibility between the first case and the fourth case, in which Plum4 suffers no outside influence (2013: 152). Vihvelin can thus reject Pereboom's second premise, which states that there is no difference between any of the cases. Similar differences exist on some readings of Pereboom's second and third cases. If the neuroscientists have designed Plum2 at birth such that he is incapable of learning, then he is not a normal agent with the ability to choose based on reasons. As such, there is a difference relevant to moral responsibility between Plum2 and Plum4. If Plum3, who was raised to be egoistic, is unable to change his character in the manner in which we assume rational adults are capable of changing, then he too lacks rational agency. Again, there is a difference between Plum3 and Plum4. Assuming that any of these details, which establish a difference relevant to moral responsibility across the set of cases, are central to the manipulation cases, Pereboom's argument fails. A soft-

line response (as opposed to a hard-line response, which would entail arguing that Plum1 is morally responsible for his actions) would succeed. But Vihvelin's objections hinge on specific readings of the details of the three cases, leaving open the charge that her response is ad hoc. Though Pereboom has not responded to Vihvelin directly, he could argue that the details on the basis of which Vihvelin rejects his argument are only accidental. Further, Vihvelin's response resembles other objections[4] to Pereboom. Because Pereboom has addressed those objections, her response requires further development.

Whereas Vihvelin's response relies on stipulating that Plum lacks relevant abilities altogether in Pereboom's early cases, I will argue that a dispositional response to the manipulation argument requires a closer examination of masks. Without a theoretical basis for what counts as a mask, one might argue that Vihvelin does not do enough to distinguish cases in which an agent is manipulated and causally determined to act in one particular way from those in which an agent merely fails to exercise her ability, or chooses not to do so. More broadly, an account of masks is necessary for any adequate dispositional account of free will. In order to determine whether an agent who did not exercise one of her abilities could have exercised that ability, dispositionalists need criteria for masks.

[4] For instance, Demetriou (2010) argues for a soft-line response on the grounds that Plum1's mental states at the time at which he chooses to kill White are not caused by his earlier mental states. Pereboom responds to her in his (2014) iteration of the manipulation argument. I discuss Pereboom's response in section 4.

3 MASKS

I will argue that manipulation arguments hinge on comparing agents whose abilities are masked[5] to agents whose abilities are not. A principled account of masked dispositions will allow us to understand why it is intervention by neuroscientists and not determinism that precludes Plum1 from being a morally responsible agent. Put simply, intervention by neuroscientists masks Plum1's rational abilities while determinism does not.

Moral responsibility is tied closely to the presence of masks. An agent whose abilities are masked ought to be considered similar to an agent who lacks the abilities in question. Just as it would be inappropriate to hold an agent who does not know how to swim responsible for failing to save a drowning child, it would be inappropriate to hold an agent whose ability to swim is masked responsible for failing to save the child.[6] Similarly, masks on one's rational abilities can render an individual not morally responsible for their actions. Consider the legal defense that an individual is not guilty by reason of insanity; the assumption underlying the practice relies on the notion that the individual could not have done otherwise, given their mental health at the time of the crime, and that they should thus not be held responsible for their action. In some cases, this

[5] Narrow abilities pertain solely to the intrinsic properties of agents. Narrow abilities in themselves bear no relationship to properties of the world around an agent, or to other properties of that agent. Narrow abilities can thus only be altered or removed, and cannot be masked. References to masked abilities should be understood exclusively as references to wide abilities.

[6] This requires the assumption that the agent had no choice in the presence of the mask. A discussion of this possibility would require a thorough discussion of tracing and is thus beyond the scope of this thesis. An agent who chooses to drink without taking the proper precautions to ensure that they will not drive intuitively seems to be responsible for any damage that they cause while drunk driving. I would hold that their responsibility can be traced to the earlier decision that they made. They are unable to choose rationally at the time of their crash, but they could have chosen otherwise when they made the decision to drink without taking the proper precautions.

might be the result of temporary insanity that would count as a mask; in other cases, an agent might be entirely unable to make rational decisions.

Because manipulated agents' abilities are masked, and because causally determined agents' abilities are not, there is a difference relevant to moral responsibility between Pereboom's (2001) first case, in which Plum1 is actively controlled by neuroscientists, and his fourth case, in which Plum4 is causally determined. But because advocates of manipulation arguments would deny that their cases are meant to depict agents whose abilities are masked, a further examination of what counts as a mask is necessary to develop this response. In this section, I will discuss features of masks that make them relevant to manipulation arguments and propose an analysis of masks that succeeds where Fara's (2008) analysis fails.

As noted above, Vihvelin's analysis focuses on the gradability of abilities and dispositions. Cases of masks must be distinguished from cases in which an object's disposition fails to manifest due to chance alone and from cases in which an object's disposition fails to manifest due to its lacking a narrow ability. Cases of masking are those in which an object's disposition fails to manifest because the circumstances in which it was exposed to the stimulus that would normally cause to its disposition to manifest (at least in a suitable proportion of cases) include some feature that prevents the disposition from manifesting as it otherwise would. A glass that falls off a table and merely happens not to break differs from a glass wrapped in bubble wrap that falls off a table of the same height. On Vihvelin's account of dispositions, the former could have broken under the exact same conditions; the latter could not have (or was far less likely to do so) because its disposition to break when dropped was masked by the bubble wrap.

Bundles of dispositions can also be masked. Consider the abilities that properly functioning cars have while driving at sixty miles per hour. An unimpeded car driving under

circumstances conducive to driving (on a highway and not an ice sheet, for instance) has the ability to slow down and stop when its brake pedal is pressed. If someone placed an object beneath the car's brake pedal, however, its ability would be masked. The object might be large enough that the brake pedal could not be pressed down at all, or small enough that it could only be pressed down slightly. The fact that objects of different sizes can mask the ability to different degrees demonstrates that masks can be graded much in the same way that abilities can. Suppose that the car in question were driving at sixty miles per hour down the highway. A car with no object under its brake pedal would be able to stop in less time than a car with such an object. How quickly a car with an object under its brake pedal would be able to stop depends on the size of the object. A car with a relatively small object beneath its brake pedal retains the ability to stop, but its ability is masked to a slight degree. A car with a relatively large object beneath its brake pedal also retains the ability to stop, but its ability is masked to a greater degree than that of the car with the small object beneath its pedal. A car with as large an object as can fit might have its ability to stop masked entirely. The same holds of a glass' disposition to break when dropped; the more bubble wrap surrounding the glass, the higher the degree to which its disposition to break when dropped is masked.

What holds here of objects' bundles of dispositions also holds of agents' abilities. Recall that we might distinguish between a baseball player's ability-to-hit-home-runs-in-batting-practice and his ability-to-hit-home-runs-in-games. Both abilities can be masked. What counts as a mask for the former might not count as a mask for the latter. The possibility of a curveball being thrown is a part of the stimulus conditions for the latter, but not typically for the former, and would thus only count as a mask on a player's ability-to-hit-home-runs-in-batting-practice.

A bright flash of light that distracted the player would count as a mask on either, as it would decrease his ability in both cases.

An analysis of masks must account for the gradability of both kinds of masks. Michael Fara (2008) has proposed the following analysis of masks:

> An agent's ability to *A* in circumstances *C* is masked iff:
> 1. The agent tries to *A*;
> 2. circumstances *C* obtain;
> 3. the agent retains the ability to *A* while trying to *A*; yet
> 4. the agent does not succeed in *A*ing (Fara 2008: 848).

I understand Fara's use of "circumstances *C*" to mirror Vihvelin's use of "opportunity." As it stands, Fara's analysis is over-inclusive in some respects and under-inclusive in others. First, Fara's analysis fails to account for the fact that abilities are gradable and thus counts some things that are not masks as masks. Consider a case in which a baseball player swings at a pitch and misses. In such a case, each condition of Fara's analysis would be satisfied; the player tried to hit a home run in the circumstances in which he is able to hit a home run, he retains his ability to hit home runs, and he fails to hit a home run. His ability to hit a home run would thus be considered masked, which, on a dispositional account of abilities, is clearly incorrect. Similarly, I am able to win the lottery in that when I buy a ticket, there is a chance that I can win. On Fara's analysis, any case in which I buy a losing ticket will yield the conclusion that my ability to win was masked, which, again, is clearly incorrect.

Second, Fara's analysis fails to account for the fact that an ability might still be masked even if the agent does not attempt to exercise it, and thus excludes some cases of masks. Consider, for instance, the following case, drawn from Ann Whittle (2010):

> Bound Ben: Ben, an excellent swimmer, has been forcibly bound to a chair. He watches helplessly as a child drowns in a lake (Whittle 2010: 10).

Because Ben realizes that he cannot exercise his ability to swim, given his circumstances, he does not try to swim. As such, his ability to swim (and thus his ability to save the drowning child) would not count as masked on Fara's analysis. But if he were to try, he would fail, which explains why he is not morally responsible for failing to save the child. In a case of a masked simple disposition, similarly, such as a glass wrapped in bubble wrap, we need not try to break the glass by dropping it to recognize that the glass' disposition to break when dropped is masked.

Third, Fara's analysis fails to account for cases in which a disposition is exercised despite being masked. Suppose that a glass wrapped in bubble wrap were dropped and chipped slightly. The glass still broke when it was dropped, but it broke to a lesser degree than it would have if it were not wrapped in bubble wrap. Even though the glass' fragility was masked, it would not count as masked on Fara's analysis because its disposition was still exercised. An analysis of masks must make use of counterfactual conditionals, in the same way that a conditional analysis of dispositions does, in order to account for cases such as Bound Ben and the glass that only chips when dropped. To say that a disposition was masked depends on the actual sequence of events, as Fara does, is not sufficient. The intuition underlying the claim that dispositions could have been exercised even if they were not supports the same approach towards masks.

To account for these problems, I propose the following analysis of masks on agential abilities:

> Some entity *M* masks an agent's ability to *A* in circumstances *C* at time *t* until some later time *t′* iff:
1. The agent retains the narrow ability to *A* until *t′*;
2. Circumstances *C* obtain at *t*;
3. If the stimulus conditions were satisfied at *t*, the agent would *A* in a lower proportion of cases or to a lesser degree than they would if *M* were not present;
4. *M* plays a causal role in (3).

The following analysis applies to simple dispositions such as fragility:

Some entity *M* masks an object's disposition to *A* in circumstances *C* at time *t* until some later time *t′* iff:

1. The object retains the disposition to *A* until *t′*;
2. Circumstances *C* obtain at *t*;
3. If the manifestation conditions were satisfied at *t*, the object would *A* in a lower proportion of cases or to a lesser degree than it would if *M* were not present;
4. *M* plays a causal role in (3).

The satisfaction of (1) and (2) depends on the context of the attribution of the ability. To return to the example of the baseball player, I take it as a default assumption that curveballs might be thrown during games, but not that they might be thrown during batting practice. The possibility of a curveball thus ought to be counted as part of the stimulus conditions for a player's ability-to-hit-home-runs-during-games, but not for a player's ability-to-hit-home-runs- during-batting-practice.[7] If a player were thrown a curveball during a game and were to miss it, we could choose to consider the even more specific ability-to-hit-home-runs-on-curveballs- during-games. There are contexts in which it will be more appropriate to examine this ability (say, when an opposing pitcher is considering how to approach an upcoming game) and contexts in which it is more appropriate to examine the broader ability-to-hit-home-runs-in-games.

The relevant circumstances should be understood in as minimal a sense as possible. In the case of the golfer, going any further than specifying that he has a club, a golf ball, and a hole would require that we adopt the strategy of getting specific, which will ultimately undermine the dispositional analysis of abilities. Though there might be cases in which the particular circumstances require that we look at a more specific disposition, a usual circumstance would not. Saying that a golfer's ability was masked by the presence of the wind is helpful if we are trying to determine whether the golfer will make his next shot. If there is little chance of wind

[7] If a pitcher does not normally throw a curveball, it might be reasonable to say that the possibility of his throwing a curveball is not part of the manifestation conditions, in which case it would be reasonable to say that the pitcher throwing a curveball would mask the batter's ability to hit a home run.

picking up again when he prepares for his next putt, then whether he is able-to-putt-successfully-when-the-wind-picks-up is less helpful than whether he is able-to-putt-successfully. Because the latter is more helpful, we should understand the wind as a mask instead of as part of the stimulus conditions of the disposition.

More must be said about what counts as a test case for a mask. Consider first how we might determine the relevant test cases for a disposition. A glass is fragile if it is such that it will break in a sufficient proportion of cases in which it is dropped on hard surfaces. Suppose that a given glass does not break on one particular drop. Such a case would count as an individual test case, but more test cases are required to determine the degree to which the glass is (or was) fragile. We must examine the same drop in a number of nearby possible worlds, changing as few details as possible about the stimulus conditions and the intrinsic property in virtue of which the object has the disposition.[8] If we were to change anything about the constitution of the glass, we would change its dispositions, and would thus no longer be testing the same disposition. By changing minor details about the air currents around the glass, the force with which the glass is knocked off of the table, and similarly relevant details, or by making slight alterations to the laws of nature, we can generate new test cases in appropriately nearby possible worlds.[9] By considering the proportion of such cases in which the glass breaks, we can determine whether the glass is fragile and how fragile it is. Which facts and laws of nature are appropriate to change

[8] We can, of course, continue to test the glass by dropping it in the actual world. It is important to note, however, that in the event that the glass should break the first time it is dropped, further tests in the actual world would be impossible. Further, there are instances of cases in which further tests in the actual world might be impossible; a presidential candidate can only run in a given election once.

[9] See Lewis (1981) for a discussion of divergence miracles, which can be stipulated to generate further test cases. See Lewis (1973, 1986: 20-27) for a discussion of the proximity of possible worlds.

depends on the context of the attribution; to paraphrase Lewis (1986: 21), discussing worlds in which glasses float when dropped will not help if we are concerned with a glass' disposition to break when dropped in the actual world.

To test whether a disposition is masked, we should take the same approach. The relevant stimulus conditions, the intrinsic facts that constitute the relevant dispositions, and the mask itself should all be held fixed. By considering the proportions of such cases in which the mask is held fixed and the disposition is not exercised, we can determine whether the object's disposition is masked and the degree to which it is.

Consider how my analysis applies to paradigmatic cases of masks:

Bubble wrap masks a glass' disposition to break when dropped at time *t* iff:

1. The glass retains its disposition to break when dropped until *t′*.
2. The circumstances in which the glass is disposed to break when dropped obtain.
3. If the glass were dropped at *t*, it would break in a lower proportion of cases than it would if it were not wrapped in bubble wrap.
4. Bubble wrap plays a causal role in the glass' not breaking when dropped.

Suppose that a particular glass, when not wrapped in bubble wrap, broke in half of the relevant test cases. When wrapped in a full layer bubble wrap, suppose that it broke in just one tenth of the relevant test cases. The full layer of bubble wrap would mask the glass' fragility to a greater degree than would a small piece of bubble wrap taped to one side of the glass. While the small piece might prevent the glass from breaking in one in each thousand cases, and would thus count as a mask, it masks the glass' fragility to a far lesser degree than does the full layer of bubble wrap. Consider how my analysis would apply to agential abilities such as Austin's (1956) golfer:

The wind picking up masks the golfer's ability to putt successfully at *t* iff:

1. The golfer retains the narrow ability to putt successfully until a later time *t′*.
2. The circumstances in which the golfer is able to putt successfully obtain at *t* (i.e. the golfer is awake, has a golf club and a ball, and is on a golf course).
3. If the golfer were to try to putt at *t*, he would miss his putt in a higher proportion of cases than he would if the wind did not pick up after he putted.
4. The wind plays a causal role in the golfer's missing his putt.

Again, each condition is satisfied, and the golfer's ability to putt successfully counts as masked on my analysis. A stronger gust of wind will mask the golfer's ability to a greater degree than a weaker gust of wind, as a stronger gust will make the golfer miss by a greater margin.

As I have argued, Fara's analysis is unable to account for specific cases: those in which an agent fails to do something because her ability is gradable, those in which an agent does not try to exercise her ability, and those in which an agent succeeds at something to a lesser degree than she would if the mask were not present. I believe that my analysis addresses each of these problems. First, it is unclear what would count as the mask in a case in which, for instance, I failed to win the lottery because I picked out the wrong ticket. Holding fixed all of the facts about my decision-making process in buying a ticket, there are nearby possible worlds in which any ticket that I bought would be the winner. Should we hold that my buying a ticket with the losing number on it counts as a mask, the conclusion that my ability to win the lottery is masked by my buying a losing ticket would be trivial. Should we hold that my buying a ticket with that particular number on it masks my ability to win the lottery, (3) would be false, because there is no reason to suppose that my particular number is any less likely to win than any other number. Such a case would thus not count as an instance of a mask; instead, I failed to win the lottery because my ability to win is incredibly unlikely to be exercised successfully. Cases in which agents fail to successfully exercise an ability because they have the ability to a low degree thus do not count as masks on my analysis.

Second, consider the difference between how my analysis would address a case like Whittle's 'Bound Ben' (2010) and how Fara's analysis would. Because Ben does not try to exercise his ability, on Fara's analysis, his ability is not masked. On my analysis, however, his

ability is masked, because condition (3) builds in a conditional that accounts for cases in which agents do not try to exercise their abilities:

The ropes tied around Ben mask his ability to save the drowning child at *t* iff:

2. Ben retains the narrow ability to save the child until *t′*;
3. The circumstances in which Ben is normally able to exercise his ability obtain (i.e. he is awake and he is near a body of water);
4. If Ben tried to exercise his ability at *t*, he would succeed in a lower proportion of cases than he would if he were not bound;
5. The ropes play a causal role in Ben's failure to save the drowning child.

Each condition is true, and Ben's ability thus counts as masked on my analysis. Finally, consider the difference between Fara's analysis and mine in whether a car with a small object beneath its brake pedal would count as having an ability masked:

The object beneath the brake pedal masks the car's ability to stop at *t* iff:

1. The car retains the narrow ability to stop until *t′*;
2. The circumstances in which the car is normally able to stop obtain (i.e. the car is driving on a road);
3. If the brake pedal were pressed at *t*, the car would slow down at a slower rate than it would if the object beneath its brake pedal were not present;
4. The object beneath the brake pedal plays a causal role in the car's decreased ability to stop.

Unlike Fara's analysis, mine accounts for the fact that the car's ability to stop is masked in this case. Further, as argued above, this analysis accounts for the fact that masks are gradable. The larger the rock placed under the car's pedal, the more its ability to stop will be masked. Drunkenness masks someone's ability to drive; the more one drinks, the more one's ability to drive is masked.

We thus have criteria for what should and should not count as a mask on a manipulated agent's ability. On a dispositional account of free will, status as a morally responsible agent depends on the absence of masks on the relevant abilities. One of the chief advantages of understanding masks as coming in degrees is that such an understanding allows for moral responsibility to come in degrees as well. There might be cases in which someone's ability to

choose is masked to a slight degree and their moral responsibility is similarly only slightly decreased. In the following section, I will argue that this analysis, when applied to Pereboom's earlier (2001) set of manipulation cases, yields the conclusion that Plum1's rational abilities are masked and that Plum4's are not, providing a principled reason to reject Pereboom's claim that there is no difference between an actively manipulated and a causally determined agent.

4 PEREBOOM'S MANIPULATION CASES

I have offered a new analysis of masks on simple dispositions and agential abilities. With this analysis in mind, we can examine Pereboom's claim that Plum1, who is actively manipulated by neuroscientists, is not morally responsible in virtue of the fact that his actions are causally determined. What holds across each iteration of Pereboom's manipulation argument is that the agent in the final case is a normal, unimpeded agent in a deterministic universe who chooses of his own accord to kill White. Such agents are not subject to masks. The mere fact that a universe is deterministic does not preclude the agent from making a different decision in nearby possible worlds. Consider my analysis applied to determinism:

> A set of deterministic laws of nature masks an agent's ability to choose on the basis of reasons at time *t* iff:
> 1. The agent retains the narrow ability to choose on the basis of reasons until some later time t_1;
> 2. The circumstances in which the agent is normally able to exercise their ability obtain at *t* (i.e. they are awake and aware of the relevant reasons);
> 3. If the agent tried to exercise their ability at *t*, they would succeed in a lower proportion of cases than they would if the laws of nature were not deterministic;
> 4. The laws' being deterministic play a causal role in the agent's failure to choose on the basis of reasons.

Assuming a dispositional account of free will, there is no reason to believe that (3) is true. As Vihvelin (2013) puts it, determinism does not matter for free will. Just as a glass that falls off a table in a deterministic universe and happens not to break could have broken, an agent in a deterministic universe who decides to act in one way could have decided to act differently. Determinism no more masks agential abilities than it masks simple dispositions such as fragility. There is no reason to believe that a set of universes with indeterministic laws of nature would produce a higher proportion of cases in which an agent succeeds at choosing to act differently than a set of universes with deterministic laws of nature would produce. One might object that the set of deterministic universes will all produce the same result, but this will only be the case if

we hold fixed the particular set of deterministic laws of nature and the past state of the universe. But if we were to do so, then it would not be determinism, but instead that particular set of laws of nature in conjunction with the past state of the universe that would be the mask, and not the mere fact that the laws of nature are deterministic. Determinism is not a mask.

In different iterations of Pereboom's argument, the first, second, and third Plums are subject to different forms of manipulation, and thus possess varying degrees of the ability to choose, understood in terms of dispositions and masks. Since first developing the argument (Pereboom 1995; 2001), Pereboom has altered his cases in light of arguably ad hoc responses to specific readings of what he would consider unimportant details. As he does in more recent iterations of the argument (Pereboom 2013; 2014), he can specify that there is a proper causal relationship between Plum1's earlier and later mental states, thus addressing Vihvelin's objection to his first case. Though he has not done so explicitly, Pereboom could also specify that Plum2 is able to learn and that Plum3 is able to change his character, addressing Vihvelin's responses to his second and third cases.

While Vihvelin's response to Pereboom suffices for those particular readings of those cases, a dispositional response should focus on whether each agent's rational abilities are masked. As I will argue below, some of Pereboom's first cases depict agents whose rational abilities are masked, and who thus do not retain their abilities to choose on the basis of reasons throughout the case. On some readings of his more recent cases, the agents' abilities are left unmasked, meaning that each retains the ability to choose on the basis of reasons during the relevant time period. For each case, if Plum's rational abilities are masked, then he is not morally responsible for his actions (or is at least less morally responsible, depending on the degree to which his abilities are masked), but in such cases there is a difference relevant to moral

responsibility between Plum and a causally determined agent whose abilities are not masked. Regarding any such cases, we should take a soft-line response. If Plum's abilities are left unaffected, then he never loses his rational abilities and remains an appropriate target of moral responsibility. Regarding those cases, we should take a hard-line response.

If Plum's ability to choose on the basis of reasons is masked, the considerations relevant to whether he is morally responsible will be the same as those for any other agent whose ability is masked. As I argue above, a person tied to a chair is not responsible for failing to save a drowning person; being tied to a chair masks their ability to do so. Their circumstances prevent them from acting as would be necessary to save the drowning person.[10] In light of this response, consider Pereboom's (2001) first case:

> Professor Plum was created by neuroscientists, who…"locally" manipulate him to undertake the process of reasoning by which his desires are brought about and modified – directly producing his every state from moment to moment. The neuroscientists manipulate him by...pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum...does not act because of an irresistible desire...and he does not think and act contrary to character since he is often manipulated to be rationally egoistic. His effective first-order desire to kill Ms. White conforms to his second-order desires. Plum's reasoning process exemplifies the various components of moderate reasons-responsiveness. He is receptive to the relevant pattern of reasons, and his reasoning process would have resulted in different choices in some situations in which the egoistic reasons were otherwise. At the same time, he is not exclusively rationally egoistic since he will typically regulate his behavior by moral reasons when the egoistic reasons are relatively weak – weaker than they are in the current situation (2001: 112-113).

[10] This is controversial; see Whittle (2010) for further discussion. Whittle argues that the relevant abilities in Frankfurt-style cases are more specific than the general abilities that all agents have, and we should hold fixed the presence of the manifestation conditions when attributing someone an ability. Jones is not "disposed-to-do- otherwise-with-the-device-present" (2010: 9). Because he lacks the ability, Whittle argues, the dispositional approach to Frankfurt fails. Though her argument is beyond the scope of this paper, I would argue that her response mischaracterizes the dispositional response to Frankfurt. Jones is, per Vihvelin's (2004) response, still disposed-to- do-otherwise-with-the-device-present until Black actually activates the device. In cases in which Black never activates the device, Jones remains able to do-otherwise-with-the-device-present.

Pereboom's argument relies on the claims that (i) Plum1 satisfies any and all proposed compatibilist conditions and that (ii) Plum1 is not morally responsible for his actions. If Plum1's ability to make rational choices is masked by the neuroscientists intervention, then (i) is false. Consider how his ability fares on my analysis of masks:

> The neuroscientists' intervention masks Plum1's ability to make rational decisions at time *t* until some later time t_1 iff:
> 1. Plum1 retains the narrow ability to make rational decisions until t_1;
> 2. The circumstances in which Plum1 is normally able to make rational decisions obtain (i.e. Plum1 is awake and consciously aware of the world around him);
> 3. If Plum1 tried to make a rational decision, he would do so in a far lower proportion of cases than he would if the neuroscientists did not intervene;
> 4. The neuroscientists' intervention plays a causal role in Plum1's failure to make a rational decision.

Each statement is true, so Plum1's ability to make rational decisions is masked. Suppose that at some moment during the process of the neuroscientists' intervention, Plum1 formed the intention to make a rational choice to refrain from killing White. Because the neuroscientists control each of his mental states on a moment-to-moment basis, the intention that Plum1 forms will not play a causal role in his later mental states. If he tried to make a rational decision, he would fail to do so in any test case in which the neuroscientists alter his mental states as Pereboom says they do. Further, because his ability to make rational decisions is masked, Pereboom's claim that Plum1's "reasoning process exemplifies the various components of moderate reasons-responsiveness" (2001: 111) is false.[11] Because Plum1's ability to make rational decisions is masked by the neuroscientists' intervention, he is not "receptive to the relevant patterns of reasons" (2001: 111). The fact that Plum1 seems to engage in a reasoning process is irrelevant; the neuroscientists are the cause of each of his mental states, and they only mimic a reasoning process.

[11] This term is drawn from Fischer & Ravizza (1998).

Though a different set of reasons might have resulted in him acting differently, such cases are irrelevant in light of the fact that the neuroscientists intervene by implanting egoistic reasons in him. As I have argued, possible worlds in which the mask is not present should not be used as test cases in determining whether an agent's abilities are actually masked, but instead should be used as a comparison against cases in which the mask is present. To say that Plum1 would act differently if the reasons available to him were different is akin to saying that a glass wrapped in bubble wrap would have broken when dropped if it were not wrapped in bubble wrap. That the glass would have broken without the bubble wrap only further supports the conclusion that the glass' disposition to break when dropped is masked, as it highlights the truth of (3). Similarly, in this case, the fact that Plum1 would have acted differently if the neuroscientists were not present and if he were exposed to different reasons only highlights that (3) is true.

Plum1's ability to choose on the basis of reasons is masked, so there is a relevant disanalogy between him and a causally determined agent whose rational abilities are not masked. As I have argued, causal determinism, unlike the kind of intervention that Pereboom describes, is not a mask. There is a clear difference between the first and fourth cases, so a soft-line response is appropriate here; the question becomes whether the line should be drawn between the first and second, second and third, or third and fourth cases. Consider Pereboom's second case:

> Plum is like an ordinary human being, except that he was created by neuroscientists, who, although they cannot control him directly, have programmed him to weigh reasons for action so that he is often but not exclusively rationally egoistic, with the result that in the circumstances in which he now finds himself, he is causally determined to undertake the moderately reasons-responsive process and to possess the set of first- and second-order desires that results in his killing Ms.White. He has the general ability to regulate his behavior by moral reasons, but in these circumstances, the egoistic reasons are very powerful, and accordingly he is causally determined to kill for these reasons. Nevertheless, he does not act because of an irresistible desire (2001: 113-114).

Pereboom specifies that "although Plum satisfies each of the compatibilist conditions, intuitively he is not morally responsible" (2001: 114) and that:

> Causal determination by factors beyond Plum's control most plausibly explains his lack of moral responsibility in the first case, and I think that we are forced to say that he is not morally responsible in the second case for the same reason (2001: 114).

The better explanation for Plum1's lack of moral responsibility is the fact that his relevant abilities are masked. Because Pereboom is wrong to conclude that Plum1 is not responsible as a result of his being causally determined by factors beyond his control, whether the argument generalizes from the first case to the second hinges on whether Plum2's rational abilities are also masked, not on whether Plum2 is also causally determined.

Pereboom writes that "in the circumstances in which [Plum2] finds himself, he is causally determined to undertake the moderately reasons-responsive process" in virtue of which he decides to kill White (2001: 114). On a dispositional account of free will, the fact that he is causally determined to reason in a particular way does not entail that he lacks the ability to choose on the basis of reasons. In this case, Plum2 tries to choose on the basis of reasons and succeeds in doing so. Nothing about this case suggests that his doing so is an aberration; in a suitable proportion of cases, he will presumably successfully exercise his ability. His ability to choose on the basis of reasons is thus not masked, and there is a relevant disanalogy between Plum1 and Plum2. Because the mask on Plum1's rational abilities is the best explanation for Plum1's lack of moral responsibility, and because Plum2's ability to choose is not masked, the disanalogy between them supports the soft-line approach I have outlined here. Plum2 is thus morally responsible for his actions. The fact that Plum1 and Plum2 are both causally determined to choose on the basis of reasons is irrelevant, as Plum1's being causally determined is not the reason that he is *not* morally responsible. On this iteration of Pereboom's manipulation argument, the line can be drawn between the first and second cases.

An advocate of the manipulation argument presented here might respond to the objection I have posed by noting that if any facts were different, the neuroscientists would have designed Plum2 differently so as to account for those facts. By the parameters of the thought experiment, there is no reason to believe that there are possible worlds in which there exist details that the neuroscientists did not take into account. As such, one might argue, if we were to change some minor detail in the environment in order to generate a new test case, we would also have to change the details about how the neuroscientists programmed Plum2. Such a response, however, would be an appeal to the finkish nature of Plum2's rational abilities.

To respond to this objection, we can appeal to the same response that Vihvelin (2004) levels against Frankfurt's cases. If Jones or his environment had been slightly different, Jones might have decided (or begun to decide) against killing White. If Jones had begun to decide against killing White, Black would have intervened so as to ensure that Jones would decide to kill White. But the fact that Black *would have* intervened does not mean that Jones *actually* loses his ability to do otherwise. In this case, the same holds true. If facts about Plum2's environment had been different, then the way in which he had been altered would not cause him to kill White in every case. But to examine test cases, we would need to hold fixed Plum2's psychological state. Though it might be true that the neuroscientists *would have* changed Plum2's programming if things were different, Plum2 was able to choose to do otherwise at the time at which he chose to kill White in virtue of what his psychological state *actually* was.

5 PEREBOOM'S NEW CASES

Pereboom's goal is to present a manipulated agent no different regarding any compatibilist capacity from a causally determined agent—an agent who is reasons responsive, has second-order desires that conform to his first-order desires, and has any other capacity that a compatibilist might say is necessary for free will (2014: 75). Because Vihvelin holds that the wide ability to choose on the basis of reasons is necessary for free will, and because retaining the ability requires that the ability is not masked, Pereboom could argue that his cases still prove his conclusion if the manipulation in the first case does not mask Plum1's ability to choose while still rendering him not morally responsible. Pereboom argues that his most recent (2013, 2014) iterations address this challenge directly by addressing a common objection to his argument—that Plum1 lacks agency altogether (2014: 76).

Though Pereboom raises this objection in response to Demetriou (2010), it can also serve as an objection to the dispositional response I have proposed here. Demetriou argues that the kind of intervention that would be necessary for Plum1 to act as the neuroscientists desire regardless of his earlier mental states would preclude the possibility of his having the kind of control over his own actions that would be necessary for him to be an agent at all. While Pereboom can alter the case so as to reduce the neuroscientists' control over Plum1, a diminished degree of control on the neuroscientists' part over Plum1's actions would suggest that Plum1 is responsible for his actions. Demetriou calls this the 'causal control dilemma' (2010: 601).

Pereboom (2014) objects to Demetriou on the grounds that cases can be drawn up so as to avoid this problem. He presents the following revised 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 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. (2014: 76-77)

As he does in his (2001) iteration of the argument, 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 only one mental state, Plum1 still exercises his rational abilities, and it is in virtue of a decision that he makes that he chooses to kill White. If faced with my objections to his (2001) iteration of the manipulation argument, Pereboom could argue that this case does not depict an agent whose rational abilities are masked. Instead, it depicts an agent whose abilities are exercised differently than they would have been without intervention. But, he would argue, there is still reason to believe that Plum1 is not morally responsible, so the best explanation for his not being morally responsible must be that he is causally determined.

Similarly, Pereboom argues that Seth Shabo's (2010) manipulation case, the Ego Button, avoids Demetriou's causal control dilemma. In it, a politician named Natasha deliberates about whether to release damaging information about an opponent. Neuroscientists who want the opponent eliminated cause a one-time shift in Natasha's reasoning process such that she becomes more egoistic. Shabo explains:

> By ramping up activity in one region of Natasha's brain while suppressing it in another, the Ego Button ensures that her reasoning about the situation will be structured around the question, 'Which of my options will best further my interests?' (2010: 376).

Shabo specifies that Natasha's "agential capacities are in no way impaired when she acts" (2010: 377). Though he does not discuss the case in terms of masks, his comment suggests that he, like Pereboom, means to depict a case in which the manipulated agent's abilities are not masked. Rather than the sort of moment-to-moment control that would suffice for a mask, Shabo seems to present a one-time intervention. If the neuroscientists merely shift the focus of Natasha's reasoning to a different question, Natasha seems to retain and exercise her rational abilities while making her decision. Shabo argues that any charge (such as Demetriou's) that Natasha lacks agential abilities fails; she is clearly engaged in a reasoning process and has the sort of first-person perspective that Plum1—whose mental states are controlled on a moment-to-moment basis—lacks (2010: 377). Natasha is not morally responsible, Shabo argues, yet, her abilities seem to be left unmasked.[12]

In both cases, the agents intuitively seem to be in full control of their mental states and their actions at the time at which they act. In both cases, the agents were manipulated in such a way that they intuitively seem not to be morally responsible for their actions. If Pereboom is right, then these cases manage to avoid my response to Pereboom's (2001) case—and Demetriou's causal control dilemma—because the agents are not morally responsible but their abilities to choose are not masked.

But recall that analyzing whether an agent's ability is masked requires a temporal index, as does analyzing whether an agent has an ability. In both cases, we can choose whether to examine the time period from just before the neuroscientists intervene or from just before the

[12] Though Shabo does not put it in terms of masks and abilities, this is how his conclusion would map onto what I have discussed.

agent begins to deliberate about their action. Suppose the following time stamps for Pereboom's (2014) first case:

t_1: The moment just before the neuroscientists press the button.
t_2: The moment at which Plum1 begins to reason.
t_3: The moment just after Plum1 kills White.

Whether Plum1's ability is masked depends on whether we examine the period from t_1 to t_3 or the period from t_2 to t_3. Consider first the time from t_1 to t_3:

> The neuroscientists' intervention masks Plum1's ability to make rational decisions at t_1 iff:
> 1. Plum1 retains the ability to make rational decisions until t_3;
> 2. The circumstances in which Plum1 is normally able to make rational decisions obtain (i.e. Plum1 is awake and consciously aware of the world around him);
> 3. If Plum1 tried to make a rational decision at t_1, he would do so in a far lower proportion of cases than he would if the neuroscientists did not intervene;
> 4. The neuroscientists' intervention plays a causal role in Plum1's failure to make a rational decision.

If Plum1 tried to make a decision at t_1, he would fail to do so. Given that the neuroscientists will intervene in the case, any decision that Plum1 makes or begins to make at t_1 will lack causal efficacy on his actions later. While some of his mental states at t_1 (his perceptions, his memories, and so on) will still play a role in his decision at a later time, he is unable at t_1 to act in such a way that he will refrain from killing White. The neuroscientists' intervention prevents him from being able to make a rational decision and from being able to do other than kill White. During this time period, Plum1's ability is masked, and he is thus not responsible for killing White.

In examining his abilities from t_2 to t_3, however, we will reach a different conclusion. On that latter time-slice, no force or entity that would count as a mask exists, as the neuroscientists do not act on him in any way during that particular time period. Pereboom might object that the intervention counts as that force, but examining the time period from t_2 to t_3 commits us to ignoring the intervention; if we should only focus on the time after the intervention occurs, then, in principle, we cannot examine Plum1's causal history. So while it is true that Plum1 is in full

control of his abilities during this time period, he is only in full control because there is no mask present. Examining only that time period, on a dispositional account of free will, Plum1 is able to do otherwise.

For the claim that Plum1's abilities are not masked to be true, Pereboom must commit to examining only the time period after the intervention, and must ignore all facts about Plum1 prior to the intervention. But it is only because of Plum1's prior mental states that we believe that he would have decided differently if there were no intervention. It is because we index his later mental states to his earlier mental states that he seems not to be morally responsible for his actions. Including any fact about Plum1's abilities prior to the intervention, which is necessary for Pereboom's claim that he is not morally responsible, would force Pereboom to widen the time period during which we determine if he has the ability such that it would include that earlier time period. Doing so shows that Plum1's rational abilities are masked. Should we choose, however, to assess Plum1's abilities from t_2 to t_3, the criteria for a masked ability are left unsatisfied, and Plum1 can be held morally responsible for his actions. Like the golfer who missed his putt but could have made it, because no mask is present, on a dispositional analysis of abilities, Plum1 could have chosen to refrain from killing White. While Pereboom is right that there is no difference in regard to moral responsibility between Plum1 and a causally determined agent, he has sacrificed his theoretical basis for the intuition that Plum1 is not morally responsible.

The same holds true of Shabo's case. Determining whether Natasha is able to choose on the basis of reasons requires that we examine a particular time period. Because of the significance of the intervention here, examining different time periods will lead to different conclusions. On some, Natasha's abilities will be masked; on others, her abilities will be left

unmasked. Let t_1 stand for a moment just before the neuroscientists press the Ego button, let t_2 stand for the moment at which Natasha begins to deliberate, and let t_3 stand for the moment at which Natasha decides to release the information. From t_2 until t_3, Natasha is able to choose on the basis of reasons—her abilities are such that if she chooses to refrain from releasing the information, she will do so.

Examining the cases in this way explains why our intuitions tell us that Plum1 and Natasha are not morally responsible but are in complete control of their decision-making processes at the time at which they act. Pereboom and Shabo draw on our intuitions from one temporal index to argue that they are not morally responsible, and from the other temporal index to argue that they are in control. But there is no theoretical reason to mix-and-match our intuitions about agency and responsibility; our intuitions about agency from each temporal index should go with our intuitions about responsibility from each temporal index. When we include the manipulation in the temporal index, Plum1 and Natasha are not responsible for their actions, but their rational abilities are masked. When we exclude the manipulation, their rational abilities are not masked, but the fact that we are screening off the manipulation from view means that there is no reason to hold that they are not morally responsible. There is no reason to match our intuitions about responsibility from the earlier index with our intuitions about agency from the later index; doing so would be similar to matching our responsibility intuitions from the later index to our agency intuitions from the earlier index.

Note that I do not commit to examining one time period or the other in order to determine whether Plum1 and Natasha are morally responsible. This is a conditional response; if you believe that the only time period we should consider when assessing whether an agent is morally responsible is the moment that the agent makes the decision, then we can assess the time period

only after the intervention occurs, but in doing so, we relinquish any theoretical backing for the intuition that either is *not* morally responsible. If you believe that we should assess the entire period, there is theoretical backing for the intuition that either is not morally responsible, but their abilities are masked. Either approach is reasonable, but there is no plausible approach to moral responsibility on which an assessment will yield the conclusion that their abilities are not masked *and* that they are not morally responsible. My claim here is only that whichever time period ultimately proves to be the more appropriate one to assess for moral responsibility will render true one and only one of the two premises of the manipulation argument. If we examine Plum1/Natasha-from-t_1-to-t_3, we should take a soft-line response. If we examine Plum1/Natasha-from-t_2-to-t_3, we should take a hard-line response. While a hard-line response might seem unintuitive, it must be emphasized that examining the later time period requires that we ignore the manipulation entirely, as considering the manipulation at all implicitly commits us to including the time at which it occurs in our assessment of agency.

6 THE ZYGOTE ARGUMENT

Mele (2006) has developed a manipulation argument that he takes to overcome the fundamental problems with Pereboom's argument. Diana, an all-powerful goddess in a deterministic universe, desires that some event should occur thirty years from now. She creates a zygote in Mary which will develop into a man, Ernie, who is causally determined to act in the way that Diana desires. Thirty years later, Ernie acts as Diana designed him to act. Despite the fact that he satisfies any and all compatibilist conditions for agency, he is not morally responsible for his actions.

Mele compares Ernie to Bernie, another agent in a causally determined universe, who, like Ernie, is causally determined to commit some morally reprehensible act in the future. But Bernie's zygote develops as a result of chance, not intentional design. Mele argues that there is no difference relevant to moral responsibility between Ernie and Bernie. His manipulation argument follows:

1. Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.
2. Concerning free action and the moral responsibility of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.
3. So determinism precludes free action and moral responsibility (Mele 2006: 189).

Say then that t_1 is the moment before which Diana intervenes and creates the embryo, t_2 is a moment before Ernie begins to deliberate about his action, and t_3 is a moment just after he acts. Should we hold that the relevant time period is t_1 to t_3, then Diana's intervention seems relevant to the question of whether Ernie is able to do otherwise and whether he is able to make rational decisions. Should we hold that the relevant time period is t_2 to t_3, then Diana's intervention seems

irrelevant. In this regard, Mele's case resembles Pereboom's (2014) case and Shabo's (2010) case.

But Diana's intervention differs in two key ways from the kind of manipulation that Pereboom and Shabo describe. First, the intervention occurs much longer ago than the manipulation that Pereboom and Shabo describe. In this sense, the case is more similar to Pereboom's second case, in which the neuroscientists program Plum2 at birth to be sufficiently egoistic as to be the kind of person who will kill White. Because Pereboom and Shabo both argue that there is no principled difference between manipulation thirty seconds, thirty minutes, or thirty years before an action occurs, this difference can be overlooked.

The second difference, however, is more significant. Pereboom and Shabo both describe manipulation of mental states. Even in Pereboom's second case, the neuroscientists alter Plum2's mental states when he is born. Mele, on the other hand, describes the creation of a zygote, not the alteration of mental states. As such, my discussion of masks seems irrelevant to Mele's argument. Diana's intervention does not seem sufficient to rob Ernie of his ability to choose on the basis of reasons, and thus does not seem to be a mask. On my account, this would entail that Ernie is able to choose and is morally responsible for his decision.

Determining whether an agent is able to do something requires that we look at a number of similar cases. We can no more determine from this one case alone that Ernie is unable to choose on the basis of reasons, understood as a bundle of dispositions, than we can conclude that a match could not have lit from a single failed strike, or that a golfer is unable to putt from a single miss. Call the world in which Ernie chooses to exercise his rational capacities and acts immorally w_1. If we hold fixed every fact about Ernie in w_1, then in a nearby possible world w_2 in which some facts or laws are slightly different, Ernie might decide to do otherwise.

Advocates of the zygote argument would likely respond by pointing out that if some fact or law were different, then Diana would have designed Ernie differently. While that is true, that does not mean that the Ernie in w_1 is unable to choose to do otherwise. If we attribute Ernie the ability to choose, as Mele does in premise (2), we must commit to holding fixed the facts about his dispositions in any test cases.

Any appeal to the fact that Diana would have designed Ernie differently if the facts had been different makes the same mistake that Frankfurt makes in arguing that Jones was unable to do otherwise because Black would have changed facts about his psychology had he chosen (or had he begun to choose) to do otherwise. The fact that his abilities would have been altered if things were different merely means that his abilities are finkish, not that he lacked the ability to choose. A dispositional account of free will thus provides grounds for taking a hard-line response towards Mele's zygote argument. On a dispositional account of free will, the fact that Ernie is able to make decisions entails that he could have chosen to do otherwise, even if Diana would have acted differently had things been different. Premise (1) is false, counterintuitive as that might sound, for the same reason that Frankfurt's argument fails.

7 CONCLUSIONS

I have argued that a dispositional account of free will is best able to defeat manipulation arguments by focusing on the fact that some manipulation cases depict agents whose rational abilities are masked. The presence of a mask is a better explanation for a manipulated agent's lack of moral responsibility than causal determinism is. In such cases, there is a difference relevant to moral responsibility between the manipulated agent and the causally determined agent. Should cases be altered such that the manipulated agent's abilities are not masked, the cases will lose any theoretical ground for the intuition that the manipulated agent is not morally responsible. Some cases require that we decide whether to assess a time period that includes the manipulation itself, but the fact that analyzing abilities and masks requires a temporal index means that whether a wider time period is appropriate has no bearing on the success of a dispositional account of free will.

The account of masks that I have developed here can be used to strengthen a dispositional account of free will beyond simply allowing for a response to manipulation arguments. Though it departs in some respects from the orthodoxy on masks, in that it requires an account of masks as gradable, such a change is necessary on an account of dispositions according to which dispositions are gradable. A principled account of masks such as mine will allow for a stronger form of dispositional compatibilism; with an understanding of the factors that should be held fixed across possible worlds, we can more easily identify the cases in which agents were able to do otherwise. Beyond the free will debate, masks are a significant problem; they are central to questions about the nature of gender (see, for instance, McKitrick 2015 and Dembroff

forthcoming). Further, questions about bias in reasoning and whether one can be morally responsible for one's bias are heavily impacted by the nature of masks. There are also potential implications for the law; as noted above, the defense of not guilty by reason of insanity depends on an intuitive concept of masks.

5. REFERENCES

Austin, JL. 1956. Ifs and cans. *Proceedings of the British Academy* 42:109-132.

Bird, Alexander. 1998. Dispositions and antidotes. *Philosophical Quarterly* 48 (191):227-234.

Clarke, Randolph. 2009. Dispositions, Abilities to Act, and Free Will: The New Dispositionalism. *Mind* 118 (470):323-351.

Deery, Oisín & Nahmias, Eddy. 2017. Defeating Manipulation Arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies* 174 (5):1255-1276.

Dembroff, Robin. Forthcoming. Beyond Binary: Genderqueer as Critical Gender Kind. *Philosophers' Imprint*.

Demetriou, Kristin. 2010. The Soft-Line Solution to Pereboom's Four-Case Argument. *Australasian Journal of Philosophy* 88 (4):595-617.

Fara, Michael. 2005. Dispositions and habituals. *Noûs* 39 (1):43–82.

— 2008. Masked Abilities and Compatibilism. *Mind* 117 (468):843-865.

Frankfurt, Harry. 1969. Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66 (23):829-839.

Fischer, John Martin. 2014. Review of Derk Pereboom's *Free Will, Agency, and Meaning in Life*. *Science, Religion, and Culture* 1 (3):202-208.

Fischer, John Martin & Ravizza, Mark. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Janzen, Greg. 2016. 'Brain-Malfunction' Cases and the Dispositionalist Reply to Frankfurt's Attack on PAP, *Australasian Journal of Philosophy*, 94 (4):646-657.

Johnston, Mark. 1992. How to speak of the colors. *Philosophical Studies* 68 (3): 221-263.

Lewis, David K. 1973. Causation. *Journal of Philosophy* 70 (17):556-567.

— 1981. Are we free to break the laws? *Theoria* 47 (3):113-121.

— 1986. *On the Plurality of Worlds*. Wiley-Blackwell.

— 1997. Finkish dispositions. *Philosophical Quarterly* 47 (187):143-158.

Martin, C. B. 1994. Dispositions and conditionals. *Philosophical Quarterly* 44 (174):1-8.

Manley, David & Wasserman, Ryan. 2008. On linking dispositions and conditionals. *Mind* 117 (465):59-84.

McKenna, Michael. 2008. A Hard-line Reply to Pereboom's Four-Case Manipulation Argument. *Philosophy and Phenomenological Research* 77 (1):142-159.

— 2014. Resisting the Manipulation Argument: A Hard-Liner Takes It on the Chin. *Philosophy and Phenomenological Research* 89 (2):467-484.

McKitrick, Jennifer. 2015. A dispositional account of gender. *Philosophical Studies* 172 (10):2575-2589.

Mele, Alfred. 2006. *Free Will and Luck*. OUP.

— 2013. Manipulation, Moral Responsibility, and Bullet Biting. *The Journal of Ethics* 17 (3):167-184.

Mumford, Stephen. 1998. *Dispositions*. Clarendon Press.

Pereboom, Derk. 1995. Determinism al dente. *Noûs* 29 (1):21-45.

— 2001. *Living Without Free Will*. Cambridge University Press.

— 2013. Optimistic skepticism about free will. In Paul Russell & Oisin Deery (eds.), *The Philosophy of Free Will: Essential Readings From the Contemporary Debates*. OUP USA.

— 2014. *Free Will, Agency, and Meaning in Life*. OUP.

Shabo, Seth. 2010. Uncompromising source incompatibilism. *Philosophy and Phenomenological Research* 80 (2):349-383.

Smith, Michael. 2003. Rational capacities. In Sarah Stroud and Christine Tappolet (eds.) *Weakness of Will and Practical Irrationality*. OUP.

Vihvelin, Kadri. 2004. Free Will Demystified: A Dispositional Account. *Philosophical Topics* 32 (1/2):427-450.

— 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. OUP USA.

Whittle, Ann. 2010. Dispositional Abilities. *Philosophers' Imprint* 10.