FREE WILL FUNDAMENTALS:
AGENCY, DETERMINISM, AND (IN)COMPATIBILITY

by

KRISTIN MARIE DEMETRIOU (née MICKELSON)

B.A., University of Wisconsin - Madison, 2001

M.A., University of Colorado – Boulder, 2006

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Philosophy

2012

This thesis entitled:
Free Will Fundamentals: Agency, Determinism, and (In)compatibility
written by Kristin Marie Demetriou (née Mickelson)
has been approved for the Department of Philosophy

Professor Robert Hanna, Chair

Professor Michael Huemer

Date: April 16 , 2012

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline.

ABSTRACT

Demetriou (née Mickelson), Kristin Marie (Ph.D., Philosophy)

Free Will Fundamentals: Agency, Determinism, and (In)compatibility

Thesis directed by Professor Robert Hanna

The concepts of agency, determinism, compatibility and incompatibility are the stock-in-trade of the free will debate. Stifling debate, however, are commonplace mistakes and oversights related to each of these key concepts. In this dissertation, I focus my attention on three serious but widely unrecognized misunderstandings/mischaracterizations related to each of these key concepts. By identifying and resolving these fundamental problems in the contemporary literature on free will, I hope to open the door for greater progress towards the resolution of one of philosophy's oldest debates, what I call "The Primary Free-Will (In)compatibility Debate".

AKNOWLEDGMENTS

I would like to thank my professors at the University of Colorado-Boulder, especially Bob Hanna, Rob Rupert, Mitzi Lee, Michael Huemer, and Robert Pasnau, for their professional and personal guidance and encouragement during the completing of this dissertation. I would also like to extend a special thanks to Bob Hanna, Rob Rupert, Michael Huemer, Graeme Forbes and Benjamin Hale for their willingness to serve on my Dissertation Committee. I also owe a deep debt of gratitude to Michael McKenna for his good-humored support over the years and for his helpful comments on early drafts of essays which appear in this dissertation. Above all, I thank my lucky stars for having worked with my advisor (and friend) Bob Hanna, whose breadth of philosophical knowledge and creative approach to solving philosophical problems continues to inspire me.

I would also like to thank a variety of other people whose generosity, in one form or another, helped me to complete this project. I am grateful to Debra Luetmer, Laura Katovich, Emma Bennett, Joyce Cash, and my mother for taking wonderful care of my little ones while I was at work. I would also like to thank Kim for sustaining me with a regular supply of the world’s greatest pad thai, my soul food.

When it comes to the support of my family, words give out. I am thankful to my mom for her resilience and optimism, which is a constant source of strength. And to Dan, a genuine philosopher and my greatest source of inspiration, I cannot begin to thank you for all the hard hours you put in so that I could get this thing done.

CONTENTS

CHAPTER

FIGURES

FIGURE

CHAPTER ONE

INTRODUCTION

1. Statement of Purpose

There are many fundamental questions which philosophers working on the free will attempt to answer. However, most of the recognizable positions in the free will literature are those which forward a solution to one or more of the following four debates:

1. The Existential Free Will Debate: Do free agents exist?
2. The (In)determinism Debate: Is determinism or indeterminism true?
3. The Primary (In)compatibility Debate: Is determinism compatible with free will?
4. The Secondary (In)compatibility Debate: Is indeterminism compatible with free will?

Now, my purpose in this dissertation is not to forward or defend a particular solution to any of these debates. Rather, my goal is to clarify each of these questions.

Towards that end, I offer four freestanding but closely related essays, each of which addresses a common misunderstanding or oversight related to one of the four debates describes above. In the first essay, “The Soft-Line Solution to Pereboom’s Four-Case Argument”, I argue that free agency requires more than a phenomenal experience of freedom. In the second essay, “Redefining Determinism”, I critique the orthodox working definition of ‘determinism’ and forward a slightly amended working definition in its place. In the third and fourth essays, “A Critique of Vihvelin’s ‘Three-fold Classification’” and “(In)compatibility”, I investigate the notions of *compatibility* and *incompatibility* and argue that philosophers have been working with an impoverished understanding of these key concepts, and hence, with an incomplete understanding of some of the most familiar views in the free will literature, e.g., compatibilism and incompatibilism.

2. Chapter Summaries

In Chapter One, “The Soft-line Solution to Pereboom’s Four-Case Argument”, I offer a critique of Derk Pereboom’s “Four-Case Argument”, one of the most famous and resilient manipulation arguments against compatibilism. I contend that the Four-Case Argument draws its power from an ambiguity in the description of the causal relations found in the argument’s foundational case. I expose this crucial ambiguity and suggest that a dilemma faces anyone hoping to resolve it. After a thorough search for an interpretation which avoids both horns of this dilemma, I conclude that none is available. Rather, every metaphysically coherent interpretation invites either a hard- or soft-line reply to Pereboom’s argument. I then consider a recharacterization of the dilemma that seems to clear the way for the defense of a revised Four-Case Argument. I address this rejoinder by identifying a still more fundamental problem shared by all viable interpretations of the manipulation cases, showing that each involves a type of manipulation which undermines the victim’s agency. Because this diagnosis supports a soft-line reply to every viable interpretation of the argument and can be endorsed by any compatibilist, I consider it the final piece of the Soft-line Solution to the Four-Case Argument. Finally, I suggest a new taxonomy of manipulation arguments, arguing that none that employs the suppressive variety of manipulation found in Pereboom’s argument offers a threat to compatibilism.[1] I revisit the Four-Case Argument in Chapter Four, where I demonstrate that Pereboom’s argument—even if it were sound—would not be an argument for *incompatibilism*.

[1] This paper appears in *Australasian Journal of Philosophy* (2010) 88.4:595-617.

In Chapter Two, “Redefining Determinism”, I discuss and criticize the orthodox working definition of ‘determinism’ in the free will debate, especially Peter van Inwagen’s famous formal expressions of the doctrine. In the first half of this paper, I argue that the assumption that there can be only one set of natural laws in the world—an assumption that van Inwagen and many others usually make—can no longer be taken for granted. According to some contemporary cosmologists, the physical world may be a “multiverse”, a world full of distinct universes just like our own, each with its own distinct set of natural laws. While such views are not widely accepted by cosmologists, the types of multiverses described in such theories do seem metaphysically possible. In addition, many philosophers (including van Inwagen) accept the metaphysical possibility of miracles. I appeal to the metaphysical possibility of miracles and multiverses to reach the conclusion that events which are “determined” by the natural laws may not occur. This leads to a new analysis of determinism, and I point to two ongoing debates in the free will literature that must proceed differently in its wake.

In Chapter Three, “Beyond the ‘Three-Fold Classification’”, I discuss Kadri Vihvelin’s attempt to define and characterize the logical relationships between free will compatibilism, incompatibilism, and impossibilism (Vihvelin 2011, 2008). I argue that Vihvelin’s definitions of ‘compatibilism’ and ‘incompatibilism’ are each flawed—the former is, at best, incomplete and that the latter is subject to counterexample. I also argue that Vihvelin’s Three-fold Classification does not correctly represent the relationship between incompatibilism and impossibilism (notable, as Vihvelin’s central goal is to articulate the relationship between these two views). I then present a better way of characterizing these three views. As part of this project, I suggest how one might close the (apparent) logical gap between “arguments against compatibilism” and “arguments for incompatibilism”.

In Chapter Four, "(In)compatibility", I present my preferred characterization of Compatibilism, Incompatibilism, and Impossibilism. I focus centrally on making sense of the imprecise notions of "compatibility" and "incompatibility". Compatibilism is often (mis)understood as a view that I call "Compossibilism", roughly the view that there is some possible world at which the thesis of determinism is true and so is the thesis that some free human-like being exists. The mere denial of Compossibilism is "Incompossibilism", a view which would be endorsed by all Impossibilists, i.e. those who deny the metaphysical possibility or logical coherence of *free will*—even those who *deny* that deterministic laws pose a threat to the existence of free agents. Thus, as Vihvelin discusses, when one defines 'compatibilism' as Compossibilism (as Vihvelin does), one cannot plausibly define 'Incompatibilism' as the mere denial of Compatibilism.

While all Incompatibilists endorse Incompossibilism, the Incompossibilist need not endorse Incompatibilism. I show that a proper definition of 'Incompatibilism' is one which expresses both the incompatibilist's uniquely incompatibilist *justification* for his modal commitments. However, this is not the end of the matter, for I argue that, contrary to popular belief, compatibilism is not equivalent to mere compossibilism. Mere compossibilism, I argue, does not express adequately the compatibilist's view that determinism is *in no way whatsoever* a threat to free will. I argue that compatibilism should be understood as a strict compatibility thesis, what I call "Strict Compatibility Compatibilism", just as incompatibilism is understood in terms of a strict incompatibility thesis, what I call "Strict Incompatibility Incompatibilism". The Strict Compatibility and Strict Incompatibility theses are (assuming one neutral background assumption) also contradictory views. Not only do I think that compatibilists must endorse Strict Compatibility Compatibilism, I demonstrate that it is reasonable to think that they already do.

Thus, I conclude that most practicing compatibilists are “Compossibility-Compatibilists”. However, because compossibilism and compatibilism are logically independent views, the compatibilist (contrary to popular belief) need not endorse compossibilism.

CHAPTER TWO

THE SOFT-LINE SOLUTION TO PEREBOOM'S FOUR-CASE ARGUMENT

1. Introduction

For over a decade, compatibilists have struggled to respond to a powerful manipulation argument developed by Derk Pereboom: the notorious "Four-Case Argument".[1] Like other manipulation arguments, Pereboom's is designed to refute compatibilism by pointing to a fundamental similarity between the effects of freedom- and responsibility-undermining manipulation and the effects of causal determinism. In the first stage of the argument, Pereboom attempts to show that an individual can satisfy a collection of the most famous compatibilist conditions for free will without satisfying the control requirements of moral responsibility. Using this strategy, Pereboom hopes to reveal that compatibilists have failed to capture even the minimal type of meaningful freedom—the type of freedom required for moral responsibility. While many other manipulation arguments stop there, Pereboom goes one step further, generating the remarkable power of the Four-Case Argument with the diagnosis that his manipulation victims lack the requisite amount of control for moral responsibility because their thoughts and behaviors are *causally determined* by their manipulators. Clearly, if this evaluation of the responsibility-undermining feature of the manipulation is correct, then the same

[1] The original version of the argument is presented in Pereboom's "Determinism al Dente" in *Noûs*, 1995. However, it is the now standard version developed in his 2002 book, *Living Without Free Will*, that will be addressed in this paper.

responsibility-undermining feature is present in *every* action performed in a causally deterministic world. Thus, the Four-Case Argument not only threatens to discredit all known accounts of compatibilism, but also aspires to show that compatibilism is *in principle* a metaphysically untenable position.

In this paper I argue that it is Pereboom's manipulation argument, and not compatibilism, that is untenable. I begin with a review of the Four-Case Argument, followed by a discussion of Michael McKenna's valuable distinction between 'hard-line' and 'soft-line' replies to arguments of this kind. I quickly depart from McKenna's treatment of the 4-CA, however, because his preferred hard-line reply fails to address many plausible, and arguably the most charitable, interpretations of the argument. More than one relevant interpretation of the argument is available, I claim, because there is an important ambiguity in the description of the causal relations found in the argument's foundational case, Case 1. In an effort to resolve this ambiguity and, thereby, make a final evaluation of the 4-CA possible, I employ my endeavor to identify all of the metaphysically coherent resolutions of this ambiguity. For each interpretation I present, I argue that it falls under one of the two horns of a dilemma. The upshot of the dilemma, I contend, is that for every possible interpretation of the 4-CA, the compatibilist is able to provide either a compelling hard-line or soft-line response to it. Since there is no interpretation of the 4-CA which cannot be answered, I conclude that the 4-CA's general attack on compatibilism fails.

In the next section, I consider a plausible alternative characterization of the dilemma which seems, at first blush, to breathe new life in the deflated 4-CA. In light of this recharacterization, it seems as though all of the hermeneutically viable interpretations of Case 1 support the 4-CA's generalization strategy, meaning that the 4-CA can still be used to show that compatibilism is in principle untenable. In response, I diagnose the root problem with all of the viable interpretations of Case 1, showing that each of these interpretations involves a type of manipulation which undermines the victim's agency. If my diagnosis is correct, it would mean that every viable interpretation suffers from the same basic defect and, so, would invite the same soft-line reply. When this collection of soft-line replies is taken as a whole, it becomes much more powerful than any one of its members—so powerful, in fact, that it provides the compatibilist with a *solution* to the 4-CA. Finally, I step back and present the foundations for a new taxonomy of manipulation arguments. I locate the Four-Case Argument in this taxonomy and conclude that any manipulation argument employing its type of manipulation is categorically defeated by the considerations I have offered.

2. The Design of the Four-Case Argument

To get the argument started, Pereboom collects five of the most popular "causal integrationist conditions" that have arisen out of the compatibilist camp. Summarizing each of Pereboom's descriptions into slogan form, the five conditions are constancy of character, lack of constraint by irresistible desire, proper conformity of first-order and second-order desires, the capacity to regulate one's behavior based upon a moderately reasons-responsive deliberation process, and the capacity to understand and regulate one's behavior based on moral reasons. Pereboom labels these "integrationist conditions" because each is designed to capture a type of integration between an agent's psychology and his actions necessary for an agent to have sufficient control to be a candidate for moral responsibility.[2]

[2] For a more detailed summary of the origins and details of these five conditions, see Derk Pereboom, *Living Without Free Will*, pp. 100-10. In brief, constancy of character and lack of constraint by irresistible desire are traditional compatibilist requirements from Hume, the latter also associated with A.J. Ayer. The third condition, requiring the conformity of higher and lower desires, is taken from Harry Frankfurt's famous hierarchical account of the freedom required for moral responsibility. The fourth condition, reasons-responsiveness, is based primarily on the account of compatibilist control offered by John Martin Fischer and Mark Ravizza, while the specific requirement for responsiveness and regulation by moral reasons is from Jay Wallace.

Notably, Pereboom emphasizes that the compatibilist causal integrationist conditions for freedom are not expected to be sufficient for moral responsibility entirely on their own. In other words, a compatibilist is not responsible for giving a *complete* analysis of moral responsibility. When a philosopher provides compatibilist conditions for moral responsibility, his main goal is to provide conditions that confirm the compatibility of determinism with the type of freedom or control required for moral responsibility, though there are also some further "implicitly understood (non-incompatibilist) conditions about agency, knowledge, and circumstance" that must be satisfied as well (Pereboom 2002: 111). As would be expected, Pereboom stipulates that the aforementioned set of background conditions for moral responsibility are satisfied in each of his four cases, in addition to the specific compatibilist conditions he is targeting.

In the first stage of the Four-Case Argument (hereafter, the "4-CA"), Pereboom offers two cases of manipulation that are designed to show that an agent can satisfy the compatibilist integrationist conditions and yet fail to be morally responsible for his behavior. His goal is to provide a case in which an individual is subjected to an intuitively freedom- and moral-responsibility-undermining form of manipulation but still satisfies the compatibilist integrationist conditions, which would establish that even the best and the brightest of the compatibilists have failed to provide sufficient conditions for the freedom required for moral responsibility. Having shown that the compatibilists have failed *so far*, Pereboom's argument would indicate a looming threat for any future compatibilist account of freedom: no matter what further condition a compatibilist might concoct to complete the set of sufficiency conditions, a manipulation argument is waiting in the wings to undermine it.

The first putative counterexample features an individual, Plum, who is designed by neuroscientists so as to satisfy the compatibilist integrationist conditions and yet does not seem morally responsible for his actions:

> Case 1. Professor Plum was created by neuroscientists, who can manipulate him directly through the use of radio-like technology, but he is as much like an ordinary human being as is possible, given this history. Suppose these neuroscientists "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, among other things, pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Plum is not constrained to act in the sense that he does not act because of an irresistible desire—the neuroscientists do not provide him with 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 situation 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 (2002: 113).

In this case, all of the compatibilist integrationist conditions appear to be satisfied, but the features of Plum that satisfy the five conditions have been covertly manipulated into place. The manipulation is clearly quite severe: during the period of manipulation, the neuroscientists directly cause Plum's every state—at least every state of his reasoning process—on a moment-to-moment basis.[3] Due to the nature of the manipulation, the intuitive response to Case 1 from

[3] As mentioned above, Case 1 is open to a wide variety of interpretations, which will be the focus of the next section of this paper. However, there are notable ways in which Case 1 is *not* ambiguous. For instance, Pereboom's story clearly states that the manipulation is carried out "moment by moment", i.e. over some extended period of time, which effectively rules out the possibility of viewing the manipulation as occurring all in one instant. Next, Pereboom makes it adequately clear that the states constituting Plum1's reasoning process are affected by the manipulation rather than, say, just the reasons and desires upon which he reasons. This is confirmed by Case 2, where Pereboom says that the programming—which is offered as a perfect substitute for the neuroscientists—

most compatibilists is that Plum is *not* morally responsible when he finally kills Ms. White.[4] The best explanation for this intuition, Pereboom claims, is that Plum's murderous act was beyond his control. More specifically, Pereboom argues that our assessment that Plum's behavior is beyond his control is best explained by the fact that the behavior was *causally determined* by the neuroscientists. Indeed, no other compelling explanation seems readily available.

Worried that compatibilists might argue that Plum in Case 1 (hereafter, "Plum1") is not morally responsible because of the moment-by-moment aspect of the neuroscientists' control over his behavior, Pereboom adds a time lag to the control exerted by the neuroscientists on Plum to create Case 2:

> Case 2. Plum is like an ordinary human being, except that he was created by neuroscientist, who, though 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 (2002: 113-14).

causally determines the way that Plum2 will "*weigh* reasons for action" (italics added). Finally, it must be that the neuroscientists use the radio signals to "directly" cause Plum1's brain states, since sending the radio signals to anything but his brain would be a quite *indirect* way to tamper with Plum1's state of mind. So, any manipulation story which does not involve the direct causal determination of the victim's brain states, specifically those constituting his process of reasoning, would stand in conflict with the manipulation case that Pereboom describes and, therefore, would fail to be a hermeneutically viable interpretation of Case 1. (As a general point, I believe that it is extremely important to avoid taking unwarranted liberties in interpreting manipulation cases, which means that one should carefully distinguish viable interpretations of a given case from nearby manipulation scenarios which may also be quite interesting. For further discussion of this point, see Section VIII of this paper, "The *New* 3-CA and Beyond".)

[4] John Martin Fischer is a high profile (semi-)compatibilist who rejects this intuition. In a direct response to Case 1, Fischer states that "Professor Plum, it seems to me, is not blameworthy, even though he is morally responsible" (Fischer, "Responsibility and Manipulation", *The Journal of Ethics 8*, p. 158). I believe that the arguments provided in this paper provide a way for proponents of Fischer's account of freedom to respond to the 4-CA without appealing to the controversial claim that Plum1 is morally responsible for the murder of Ms. White.

As in Case 1, the intuitive response is that Plum in Case 2 (“Plum2”) is not morally responsible for killing Ms. White because his murderous act was beyond his control, having been causally determined by the neuroscientists. Thus, despite the addition of the time lag, Pereboom’s original argument to the best explanation seems to hold. Having established that the time lag makes no difference between Case 1 and Case 2 in terms of moral responsibility, Pereboom states that Case 2 alone is a satisfactory counterexample to the sufficiency of the compatibilist integrationist conditions. Thus, if *either* Case 1 or Case 2 is successful, so is the first step of Pereboom’s argument.

In the second stage of the argument, Pereboom employs a generalization strategy, constructing a bridge case from his purported counterexample cases to the case of a normal human in a deterministic world. Given that my critique of the 4-CA focuses almost entirely on Case 1 and Case 2, I will forgo a detailed review of the final two cases. Suffice it to say, the bridge case, Case 3, is a near-normal situation in which overbearing parents impose rigorous training on young Plum. Pereboom expects that the intuitive response to Case 3 will be that Plum (“Plum3”) *is* morally responsible for murdering White, despite the rigor of his training. The trouble for compatibilists is that there seems to be no principled difference between the first two cases and Case 3 that could justify holding Plum3 responsible while denying of responsibility to Plum1 and Plum2.

Worse yet, if Pereboom is correct that the responsibility-undermining feature of Case 1 and Case 2 is the fact that the victim is unable to control his behavior because it is causally determined by the neuroscientists, then it appears that the compatibilist will be forced to admit that another Plum, one embedded in a causally deterministic world, cannot be responsible for any of his actions either! To drive home this point, Pereboom concludes his generalization argument with the presentation of a fourth Plum ("Plum4") who is embedded in a causally deterministic world. Though Plum4 satisfies the compatibilist integrationist conditions and is intuitively a free and responsible agent, the responsibility-undermining feature identified in Case 1 and Case 2 is present in Case 4, i.e. Plum4's actions are causally determined. Without a principled way to distinguish Plum4 from the other Plums, our moral assessment of Plum4 must align with our assessments of the previous three Plums. The compatibilism-refuting conclusion now seems unavoidable: Plum4 is not morally responsible for killing White.

3. Discourse on the Dialectic

In a recent article, McKenna recommends a general strategy for compatibilists wishing to respond to manipulation arguments such as the 4-CA (McKenna 2008). He suggests that the compatibilist has two options: she can pursue either a "hard-line" or "soft-line" reply. Defenders of the hard-line start by seeking out an interpretation of the manipulation which, in keeping with the spirit of Pereboom's stipulations, satisfies all of the conditions that the hard-liner considers necessary for free and responsible agency. Of course, as McKenna explains, once it is clear that the manipulation victim satisfies all of these conditions, the compatibilist can reasonably respond that Pereboom's manipulation victims are free and morally responsible after all, and thereby undermine the 4-CA.[5] By contrast, soft-liners start by accepting an interpretation of the manipulation which generates the key intuition that the manipulation is responsibility-undermining. Thus, soft-liners accept the challenge of showing how the manipulation victims *differ* from agents who are free and responsible. In order to meet this burden, it seems that the soft-liner must either (1) reveal that Pereboom's manipulation victims fail to satisfy a condition which she has previously claimed to be necessary for free and/or responsible agency, or (2)

[5] For a categorization of some of the most famous attempts to respond to the 4-CA along the soft/hard divide, see "Hard- and Soft-Line Responses to Pereboom's Four-Case Manipulation Argument" by Ishtiyaque Haji and Stefaan Cuypers, in *Acta Analytica* 21, 2006, pp. 19-35.

provide a new (but not *ad hoc*) condition which establishes a principled, freedom- or responsibility-relevant difference between the manipulation victims and individuals embedded in a deterministic world. McKenna ultimately endorses the hard-line strategy, arguing that it is impossible for the compatibilist to develop a successful soft-line reply.

Like McKenna, I believe that a hard-line must be taken in responding to the 4-CA—but I do not think that the compatibilist can take an *exclusively* hard-line. While a detailed critique of McKenna's view of the dialectic cannot be undertaken here, one of the basic mistakes underlying his conclusion is that he believes the 4-CA can be answered with a single hard-line reply. In fact, no single hard-line reply could be sufficient to answer the challenge of the 4-CA because there are multiple ways of interpreting the manipulation described in Case 1 of the argument and each requires its own response. As we shall see below, there are several interpretations which successfully neutralize the *prima facie* intuition that the manipulation victims are not morally responsible, making each a candidate for a distinct hard-line reply. However, even a collection of these hard-line replies would fail to provide an adequate response to the 4-CA because there are other interpretations of Case 1 which only serve to solidify the intuition that the manipulation robs its victim of moral responsibility. For each of these latter interpretations the challenge of the 4-CA remains: Can the compatibilist explain why the manipulation victims are not morally responsible without undermining her preferred version of compatibilism in the process? No hard-line reply can meet this challenge; only a soft-line reply will do.

4. The Causal Control Dilemma

Pereboom's description of the manipulation in the foundational case of the 4-CA, Case 1, is obviously quite vague. Of course, many of the fine details that one might add to flesh out these manipulation scenarios are of little import to the overall argument and, rightly, Pereboom does not dwell on such minutiae. However, the success of the argument does depend on there being at least one coherent way to flesh out the metaphysical details of the responsibility-undermining manipulation, and it is not obvious that this can be done. The worry arises from an odd tension between some of the key details in Case 1: on the one hand, the neuroscientists are responsible for "directly producing (Plum1's) every state", but on the other hand we are told that "(Plum1) will…regulate his behavior". The tension between these two stipulations is only increased when Pereboom explicitly states that the neuroscientists exercise *causal control* over Plum1's actions which allows them to regulate Plum1's behavior. After all, it should be uncontroversial that in order for Plum1 to "regulate" his own behavior, he too must exercise some minimal causal control over his actions. As it stands, then, Plum1 and the neuroscientists seem to be competing for causal control of Plum1's states, leaving the exact nature of the manipulation far from clear. What is clear, however, is that the success of the 4-CA cannot be properly evaluated until this crucial ambiguity is resolved. This is because Case 1 will support the 4-CA only if there is a specific account of the relation between the causal contributions of the neuroscientists and those of Plum1 in producing Plum1's act of murder which explains how it would be possible for Plum1's causal contribution to be sufficient for him to causally regulate his own behavior despite the independent causal control exerted over him by the neuroscientists.

I believe that a dilemma looms for the proponent of the 4-CA who takes up the challenge of resolving this ambiguity and providing this specific account.[6] If the proponent offers an account on which Plum1 could be said to 'win' this causal competition such that Plum1 exerts independent causal control over his own behavior, then the compatibilist can reasonably counter with a hard-line reply. That is, once it is clear that neuroscientists lack the causal power to interfere with Plum1's causal control over his own states, the (so-called) manipulation would be

[6] The general argument articulated in this section was inspired by Jaegwon Kim's "Explanatory/Causal Exclusion Problem". See, for instance, Jaegwon Kim, *Mind in a Physical World* (1998), and, "The Nonreductivist's Troubles with Mental Causation", *Supervenience and Mind* (1994).

so innocuous that there would be no remaining reason to think that Plum1 is not a candidate for moral responsibility. But suppose on the other hand that the proponent of the 4-CA adopts an interpretation on which Plum1 'loses' the causal competition. Even though this sort of interpretation would generate the key intuition that Plum1 is not responsible, it would do so in virtue of Plum1's violating the two compatibilist integrationist conditions requiring that he be able to self-regulate—all the required ingredients for a soft-line reply. In order for the 4-CA to threaten compatibilism, there must be an account of the manipulation that avoids both horns of this dilemma, what I call "The Causal Control Dilemma", by somehow granting Plum1 the causal power to regulate his own behavior while yet generating the intuition that Plum1 is not morally responsible.

The proponent might try to avoid each horn by appeal to overdetermination: perhaps *both* Plum1 and the neuroscientists exert independent and equally efficacious causal power in bringing about Plum1's each and every state. At this point, we could quickly get mired in a discussion about the viability of overdetermination, getting bogged down in the controversy over the metaphysical possibility of overdetermination in isolated instances, let alone the possibility of the pervasive overdetermination that would be required for Plum1 and the neuroscientists to overdetermine Plum1's every state. Luckily, a journey into that treacherous territory is avoidable, given that Pereboom could not successfully appeal to overdetermination to explain the causal relation between Plum1 and the neuroscientists. Given the details of Case 1, Pereboom would have to be talking about perfect state-by-state overdetermination: Plum1 and the neuroscientists never diverge in purpose, with the result that Plum1 never has a non-overdetermined state during the manipulation. So, even if we imagine (for simplicity) that the overdetermination interpretation leaves open the physical possibility that Plum1 could have attempted to do something other than what the neuroscientists caused him to do, it just so happens that he never does make such an attempt; even in the absence of the causal contributions of the neuroscientists, Plum1 would have behaved no differently. Indeed, by its very definition, the overdetermination interpretation guarantees that Plum1's causal contribution alone is *sufficient* to bring about all of the states leading up to the murder. So, even if Plum1's states are overdetermined, it would seem reasonable to conclude that Plum1 has sufficient control to be morally responsible for that murder—a hard-line reply.

Perhaps causal interactionism provides a more promising escape-route from the dilemma posed above? On this strategy, the scientists and Plum1 are each causes of Plum1's behavior in virtue of being alternating links on the same causal chain which brings about Plum1's states. McKenna seems to endorse an interpretation of this kind in mounting his hard-line response to the 4-CA, suggesting that one might consider the neuroscientists to be "causal prosthetics", transmitting causal messages between Plum1 and his environment and, presumably, between Plum1's states as well. "On this model", says McKenna, "while (the neuroscientists are) able to steer Plum in certain directions (like to kill Ms. White), often times, (the neuroscientists are) functioning merely as a sort of extra causal link in a chain. (The neuroscientists function) like a prosthetic, allowing Plum to deal with his world like any other agent" (McKenna 2008: 149-50).

Following McKenna's lead, let us consider a case in which the neuroscientists are slavish causal prosthetics who faithfully convey causal signals between Plum1's states, such that the neuroscientists cause precisely the same states in Plum1 as Plum1's antecedent states would have caused by themselves in the absence of the neuroscientists. In effect, the neuroscientists employ their causal powers in the service of Plum1, so it seems that they once again lose the competition for causal control over Plum1 states—only Plum1 truly controls or regulates his behavior. Once the compatibilist is convinced that Plum1 exercises such control of his behavior, though, it is likely that her intuition that Plum1 is not morally responsible will dissolve. Indeed, McKenna worries that this interpretation makes it *so* obvious that Plum1 is a morally responsible agent that adopting it might be seen as reducing the 4-CA to a non-starter (McKenna 2008: 150, fn. 6). Viewing the neuroscientists as 'faithful prosthetics', then, apparently leads to another compelling hard-line reply to the 4-CA.

Equally problematic, however, is the scenario in which the neuroscientists interpose themselves between Plum1's states but *fail* to act as perfectly faithful causal prosthetics, such that they cause Plum1 to behave *differently* than his prior states would have caused him to act. Admittedly, such an '*unfaithful* causal prosthetic interpretation' is in the spirit of Case 1. It fits well with claims like, "The neuroscientists manipulate (Plum1) by, among other things, pushing a series of buttons *just before he begins to reason* about his situation, thereby *causing his reasoning process to be rationally egoistic*" (Pereboom 2002: 113; italics added). This passage describes Plum1's reasoning as being causally initiated by the neuroscientists, regardless of what would have followed naturally from Plum1's prior states, so that Plum1 thinks and behaves any way the neuroscientists decide. Understanding the manipulation in this way would surely lead to the key intuition that Plum1 lacks moral responsibility for his actions, which means that it generates the intuition that the 4-CA depends upon. The problem with the unfaithful prosthetic approach, though, is that when the neuroscientists are unfaithful in conveying the causal signals between Plum1's states, the neuroscientists once again win the competition for causal control of Plum1's states. Since on this interpretation it is impossible for Plum1 to exercise causal control over his own behavior, it is not amenable to a hard-line reply. It does, however, suggest a soft-line reply based on the fact that Plum1 does not satisfy all of the compatibilist integrationist conditions. Thus, the 4-CA offers no genuine threat to compatibilism on either the faithful or unfaithful prosthetic interpretations.

Only one interpretation of the causal relations underlying the dual regulation of Plum1's behavior seems left to discuss: one wherein the neuroscientists and Plum1 compose a jointly sufficient cause for each of Plum1's states, i.e., one on which neither the scientists nor Plum1 alone is sufficient to bring about Plum1's states, and only together are they able to bring about Plum1's states.[7] Unfortunately for Pereboom, a closer look reveals that the jointly sufficient cause interpretation gives rise to a similar dilemma to the one that undermined the causal prosthetic interpretation.

Pereboom tells us that Plum1 "is as much like an ordinary human being as is possible" (2002: 113), so it seems reasonable to assume that, if the neuroscientists had simply released Plum1 into the world upon his creation and performed no further manipulation on him then Plum1 would have been able to act like an ordinary human being. This, in turn, suggests that the causal contributions of Plum1's states were designed to be sufficient to bring about his subsequent states. It seems, then, that the jointly sufficient cause interpretation could only work if the neuroscientists, as part of their manipulation of Plum1, *undermine* the causal sufficiency of Plum1's states in some way. Now, one can imagine various stories about how the neuroscientists could do this, but the details will ultimately be of little import. When the neuroscientists assert their own causal powers to jointly cause Plum1's behavior, they would have to do so in one of

[7] The reader may note that a supervenience relation has not been discussed. This is because supervenience is a non-starter in this context. Given the details of Pereboom's story, it seems that if Plum1's states and causal powers were supervenient on those of the neuroscientists, then Plum1's states and causal powers would simply reduce to those of the neuroscientists and Plum1 would clearly lack the independence to be a morally responsible agent. In order to a defend a *non-reductive* account of Plum1's supervenient causal powers, a.k.a. "strong emergentism", one would have to provide a positive metaphysical story of how it is possible for such new and independent causal powers to emerge from and then causally influence the subvenient base. At best, the proponent of this view would have to solve Kim's Exclusion Problem before appealing to this type of relation to save the 4-CA.

two ways: either the neuroscientists make Plum1 behave just as he would have behaved in their absence or they cause Plum1 to behave differently than he would have behaved in their absence. As discussed in the faithful prosthetic interpretation, when the neuroscientists use their causal powers to bring about exactly the same states in Plum1 as would have resulted in their absence, it seems reasonable for the compatibilist to believe that the neuroscientists' influence does not undermine Plum1's moral responsibility for his resulting actions—which is to say, the compatibilist can give a compelling hard-line reply. On the other hand, if the neuroscientists use their causal powers to make Plum1 act differently than he otherwise would have, the neuroscientists would thereby undermine Plum1's moral responsibility. However, if the neuroscientists change Plum1's behavior in the latter way, then the compatibilist can use the same argument used against the unfaithful prosthetic version of the interactionist interpretation discussed above. Namely, the compatibilist can offer the soft-line response that Plum1 is not morally responsible because the unfaithful changes to Plum1 made by the neuroscientists undermine his ability to regulate his own behavior—Plum1 would have done otherwise had only things been left up to him.

I hope that the gravity of the Causal Control Dilemma is now clear. On one hand, we have the interpretations of the manipulation on which Plum1 *wins* the competition for causal control of his states, retaining enough causal control that compatibilists would consider him morally responsible for his actions. On each of these interpretations, Case 1 fails to generate the intuition that is needed in order to run the 4-CA and, so, each invites a persuasive hard-line reply. On the other hand, we have the interpretations on which Plum1 *loses* the competition to the neuroscientists. Each of these latter interpretations leads to the intuitive response that the 4-CA depends on, making it necessary for the compatibilist to identify a responsibility-undermining feature of the manipulation. But we have seen that, in response to each case, the compatibilist is able to point to the same feature, which means that she can offer the same soft-line reply to each interpretation. In short, the manipulation victim fails to be morally responsible because he does not have the causal control required to self-regulate, and thus cannot satisfy the compatibilist integrationist conditions—conditions which a normal agent in a deterministic universe, like Plum4, could satisfy. Ultimately, it appears that there is no interpretation of the causal relations between Plum1 and the neuroscientists that can preserve both the stipulations and intuitions that the 4-CA depends on.

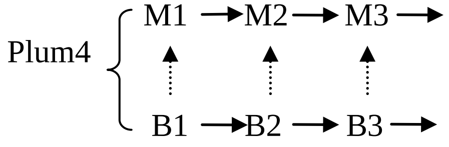
5. The Soft-Line Solution: Part One

But wait—Pereboom expects that the intuitive response to the story he tells in Case 1 will be that Plum1 is not morally responsible because Plum1's behavior is causally determined and therefore beyond his control. Reflecting on these central features of the 4-CA, one might begin to wonder if the Causal Control Dilemma is really as devastating as it appears. First of all, each of the interpretations of Case 1 falling under the first horn of the dilemma, i.e. those in which Plum1 wins the causal competition, fail to generate the expected non-responsibility intuition. The success of the 4-CA straightforwardly depends on its ability to generate this intuition, so fleshing out Case 1 in accordance with any of these interpretations would reduce the 4-CA to a non-starter. Thus, even though each is a *metaphysically coherent* interpretation of Case 1, one might reasonably argue that each of these interpretations is so horribly uncharitable to the 4-CA that none can be considered a *hermeneutically viable* interpretation—especially in light of the fact that more friendly alternatives exist. Assuming this is right, and I believe it is, the Causal Control Dilemma should be seen first and foremost as separating the unviable interpretations from the viable ones. In light of this recharacterization, it becomes clear that each of the hard-line replies discussed above are directed at unviable interpretations of the 4-CA—so, properly speaking, they are not directed at the 4-CA at all—which means that they do not indicate any weakness in Pereboom's argument. Ultimately, then, the success or failure of the 4-CA must be determined by the quality of the soft-line replies given to the viable interpretations of it.

Once we narrow our focus to the soft-line replies, though, the proponent of the 4-CA might insist that the content of these replies actually highlights the success of the most important aspect of the 4-CA: the generalization strategy. Upon review, the proponent might argue, the viable interpretations of Case 1 generate the intuition that Pereboom expects and, it seems, for precisely the reason that Pereboom identifies: Plum1 intuitively lacks the control required for moral responsibility because his actions are *causally determined*. If this is right—and the compatibilist already seems to have agreed that it is—the 4-CA still leads to a conclusion that is devastating to compatibilism: Plum4, the normal agent in a deterministic world, lacks the control required to self-regulate and so cannot be morally responsible simply because his states are causally determined. Here, then, the original generalization strategy of the 4-CA is operating in full effect, apparently establishing that compatibilism is in principle an incoherent position. Now, in order to adopt this line of defense, one must sacrifice Pereboom's claim that all of the compatibilist integrationist conditions are satisfied by Plum1, but this is hardly problematic. The proponent might easily argue that, in light of the success of its generalization strategy, the 4-CA not only shows that compatibilism is in principle false, but also that a determined agent cannot satisfy even the most anemic of the compatibilist causal integrationist conditions. It seems, then, that the 4-CA still points to an embarrassing flaw in (at least some) contemporary accounts of compatibilism while on its way to rule out all of them. This shows, one might conclude, that the 4-CA emerges virtually unscathed from the purportedly insoluble Causal Control Dilemma.

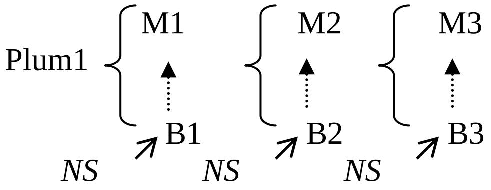
Fortunately for compatibilism, the compatibilist can block even this revitalized version of Pereboom's generalization strategy. This is because there is an important difference between, on the one hand, cases in which one's behavior is deterministically caused by such things as Pereboom's brain-tweaking manipulators, and on the other hand cases in which one's behavior is causally determined by one's own prior states (as would be the case in a causally deterministic world). While others have made similar attempts to defend the existence of a morally relevant difference between these scenarios, supporters of the 4-CA have been less than impressed because, hitherto, the metaphysical underpinnings of this difference have not been adequately exposed. However, now that we know that the only interpretations of Case 1 which generate the intuition that Plum1 is not morally responsible are also those in which the neuroscientists win the competition for causal control over Plum1's states, it is possible to expose the fundamental difference between the causal relations that obtain in the viable interpretations of Case 1 and those that obtain in Case 4. We have seen that when the neuroscientists win the competition for causal control of Plum1's states, it is because the neuroscientists unilaterally initiate changes in Plum1's states. With that in mind, consider the following diagrams illustrating the causal relations between the Plums' bodily/brain states (B), the phenomenological mental states (M) associated with (B), and the manipulative neuroscientists (NS)[8]:

1a. Plum as Normal Human Person in a Causally Deterministic World (Case 4):



[8] This style of diagram is often used in philosophy of mind to represent different visions of mental causation. I believe that my arguments are effective regardless of one's preferred theory of mental causation, so I leave it to the reader to fill in and deal with the unrelated challenges resulting from his/her views on mental causation.

2a. Plum as Causally Regulated by Neuroscientists (Case 1)[9]:



A deep difference between Plum1 and Plum4 is immediately apparent: Plum1 is not a causally integrated entity in the same way as Plum4.[10]

Along a similar line, John Martin Fischer and Mark Ravizza have suggested that an individual like Plum1 might not be a "coherent self" and this explains Plum1's lack of moral responsibility (Fischer 1998: 234-5, fn. 26). Although this response is intuitively compelling, it has been met with serious opposition. Pereboom claims that there is no reason to suppose that Plum1 is not a coherent self because "one might imagine that Plum's mental states in Case 1 or Case 2 are qualitatively identical over time to those of a non-manipulated person" (Pereboom 2002: 121). Clearly, the above diagrams lend support to Pereboom's response to Fischer and Ravizza, as they represent Plum1 and Plum4 as having the same qualitative experiences despite the differences in their circumstances. However, even if Plum1 and Plum4 have exactly similar

[9] One might wonder how information about Plum1's states is transmitted to the neuroscientists in this scenario, given that such transmission presumably would be causal and no causal route running from Plum1 to the neuroscientists is represented. To clarify, the lack of a causal arrow here simply reflects the lack of a *direct* causal relation between Plum1's bodily state and the subsequent button-pressing by the neuroscientists. That is, I do not want to deny the presence of a causal chain which could account for the neuroscientists knowledge of Plum1's every state. What this diagram is designed to show that Plum1's states are not the *proximate causes* of any of neuroscientists' button-pressings (while, on the other hand, the proximate cause of any one of Plum4's states is his own prior state). That is, Plum1's states do not, on their own, causally necessitate that the neuroscientists press the buttons that they do. Rather, they press the buttons they do as a causal result of their own, independent reasoning—meaning that the neuroscientists are free to decide, for reasons all their own, which state to cause in Plum1 at any given moment of the manipulation.

[10] This diagram will be useful even if one wishes to argue that Plum1 might have causal integration between some of his states even though his reasoning process and behavior is different than it would have been due to the causal input of the neuroscientists. In such cases, Diagram 2a could be seen as scoping down on the precise location of the failure of agency that occurs where the neuroscientists causally regulate the isolated area of Plum1's brain/body which constitutes his reasoning process. (Of course, it is now highly suspect to call the causally disjointed series of states at issue a "reasoning process".) This narrowing of scope does not affect my argument, of course, for the states constituting one's reasoning process are the most central to one's agency (at least the robust sort required for moral responsibility), and so a failure of causal integration among these states alone would be sufficient to undermine Plum1's agency.

physical and qualitative states, this does not ensure that Plum1 and Plum4 have the same status in terms of *agency*—a point that Fischer and Ravizza's reply fails to drive home.[11] With the above diagrams in hand, we can see now that even if we were to grant Pereboom's point that a unified consciousness could arise from the manipulated brain in Case 1, and even if we were to grant that this entity had sufficient unity of conscious to be a coherent self, the compatibilist still has reason to reject that this 'self' is an agent. As displayed in Diagram 2a, Plum1's physical and qualitative mental states are not causally efficacious in bringing about his subsequent physical and mental states; Plum1's states are, rather, the end effects of the causal powers expressed by the neuroscientists. I take it to be uncontroversial that when the neuroscientists suppress the causal efficacy of Plum1's states, taking the causal regulation of Plum1's states into their own hands, that they thereby suppress his agency. With that in mind, I refer to this type of manipulation as "suppressive manipulation". By contrast, a compatibilist would consider the causally integrated Plum4 depicted in Diagram 1a to be a paradigmatic agent.[12] So, by attending to previously overlooked details, the compatibilist is finally in a position to identify a problem common to all of the viable interpretations of Case 1 that she has identified, a problem that does not generalize to Case 4.

Thus, it would appear that there is a significant difference between the effects of one's being causally determined by suppressive manipulation and the effects of being an inhabitant of a causally deterministic world. What is more, since the neuroscientists undermine Plum1's

[11] The larger problem with Fischer's strategy, of course, is that it offers no clear interpretation of the manipulation cases, it provides little argument in favor of the general interpretation it assumes, and, so, leaves open the possibility of a more charitable reading of the cases which could side-step his criticisms. By contrast, the strategy in this paper is to leave *no* possible interpretation without a definitive response.

[12] More precisely, Plum4 provides an uncontroversial *base* for being a paradigmatic agent, insofar as his states are causally efficacious in bringing about his subsequent states. The remaining details of how his mental states are interrelated and related to his physical body must yet be filled in some appropriate way. However, the crucial point is that while such details feasibly can be filled in for Plum4; by contrast, neither the physical nor the mental states of Plum1 are causally efficacious in regulating Plum1's behavior.

ability to self-regulate by disrupting his agency, it is plain that the compatibilist need not appeal to any of the controversial details of the causal integrationist conditions in order to give a decisive soft-line reply to every viable interpretation of the 4-CA. Since *any* compatibilist can endorse this series of soft-line replies, it seems that there is now a soft-line *solution* to Pereboom's challenge: the principled difference between Case 1 and Case 4 is that Plum4 is a fully integrated agent but Plum1 is not (and could not be so long as the suppressive manipulation continues).[13]

6. The Soft-Line Solution: Part Two

At this point, the reflective reader might notice that there is something suspect about the *prima facie* intuitions that I attribute to the compatibilist in the previous sections. Now that we have established that two individuals with exactly the same bodily and mental states can differ with respect to agency, it is no longer obvious that Plum1 is morally responsible just because the neuroscientists faithfully bring about the states in him that would have obtained in their absence. This means that the compatibilist will not be in a position to render a final judgment about Plum1's moral responsibility in the scenarios where the neuroscientists faithfully produce Plum1's states until she knows what accounts for the fact that the neuroscientists are faithful. In other words, the same details which were used to illuminate the problem shared by the "unfaithful" versions of the prosthetic interpretation and the jointly sufficient cause interpretation also indicate that the compatibilist should revisit their "faithful" counterparts. As we shall see,

[13] *Even for most libertarians*, an individual would not qualify as an agent if he has no causal control over his choices or bodily movements. So, as long as the incompatibilist agrees that some causal contribution to one's subsequent states is required for agency (at least among the states constituting one's process of reasoning, as mentioned above in footnote 10), the incompatibilist should agree that the defense offered here is not driven by particularly compatibilist commitment—whether deterministic or indeterministic, the causal connections between Plum1's states are suppressed by the neuroscientists. This fact may be of interest to libertarians who endorse event-causal indeterminism, for Pereboom has claimed that such libertarians and compatibilists are in the same sinking boat when it comes to answering the challenge of manipulation arguments like the 4-CA (see, for example, "Living Without Free Will: The Case For Hard Incompatibilism" (in *The Oxford Handbook of Free Will*, New York: Oxford University Press, 2002), p. 478.

once these details are revealed, the compatibilist will have to reject her *prima facie* intuitive responses to these interpretations of Case 1.

Starting with the faithful prosthetic interpretation, recall that McKenna describes the neuroscientists as “functioning merely as a sort of extra causal link in a chain”, along with the use of the term ‘causal prosthetic’. This description gave McKenna’s interpretation the appearance of being an instance of the interactionist interpretation that we were looking for above. That is, it prompted us to imagine a causal chain in which Plum1’s states retain their causal efficacy. On this chain, the neuroscientists cause Plum1’s behavior insofar as they provide the proximate cause of Plum1’s thoughts and actions, but Plum1’s states are the proximate cause of the neuroscientists’ pressing the buttons they do, allowing for the judgment that Plum1’s states are causally responsible for his subsequent states. Viewing the case this way, which seems to be in the spirit of McKenna’s proposal, it appears obvious that Plum1 could be a morally responsible agent. Indeed, a diagram depicting this causal story would be relevantly similar to Diagram 1 above, showcasing Plum1 as a strange, but causally integrated agent.[14] Thus, a *genuine* interactionist interpretation of the manipulation would generate an intuitive response that would make plausible the hard-line reply McKenna that offers. Unfortunately for McKenna, though, since this version of his interpretation fails to generate the crucial non-responsibility intuition, the import of his hard-line reply is arguably undercut by the fact that it responds to an unviable interpretation of Case 1.[15]

[14] In order to construct this diagram, one would simply have to (1) redraw Diagram 1a, (2) add an ‘NS’ between B1 and B2 and again between B2 and B3, and (3) insert an arrow of causation between the latter symbols to generate a new version of the causal chain represented in Diagram 1a.

[15] To be clear, though, the critique of the 4-CA being developed in this section does not depend on whether this interpretation is in fact unviable—although I believe it is. For anyone who thinks that it *is* a hermeneutically viable interpretation, the fact that a hard-line reply to the 4-CA would be forthcoming is, all on its own, sufficient to preclude any hope of utilizing this interpretation to save the 4-CA.

However, we can also take McKenna's description of his prosthetic interpretation at face value. When we do so, there are aspects of his description which make it incompatible with a straight-forward interactionist reading. Recall that McKenna allows the neuroscientists the flexibility to "steer" Plum1 as they see fit, when they see fit. So, while they might faithfully *choose* to cause precisely the same states in Plum1 that would have obtained in their absence, the neuroscientists might just as easily choose to initiate changes in Plum1 that would not have occurred in their absence. To see why it is problematic that the neuroscientists are able to choose which states they cause in Plum1, consider a period during which the neuroscientists fail to act as perfectly faithful causal prosthetics, such that they cause Plum1 to behave *differently* than his prior states would have caused him to act. In such a case, Plum1's states are causally initiated by the neuroscientists, so that Plum1 thinks and behaves *any* way that the neuroscientists happen to decide. As shown in Diagram 2a, this sort of manipulation undercuts Plum1's moral responsibility for his actions by undermining the causal integration required for Plum1 to be an agent. However, it should now be evident that *even if the neuroscientists happen to be perfectly faithful*, causing only those states in Plum1 that would have been caused naturally in their absence, Plum1 would still lack the causal integration required for agency! So long as the neuroscientists serve as the independent proximate causes of Plum1's states, such that their button-pressings are expressions of their own desires rather than the effect of the causal powers exerted over them by Plum1's prior states, it is Diagram 2a which accurately depicts the neuroscientists causal relation to Plum1.[16] Assuming that the compatibilist should renounce their under-informed *prima facie* intuition to the interpretation McKenna describes rather than

[16] Notably, once we abandon the idea that Plum1's states *cause* the neuroscientists to press the buttons which then *cause* his subsequent states, then it no longer seems appropriate to say that Plum1's states and the actions of the neuroscientists are alternating links on the same causal chain. Of course, if there is no causal chain, then the neuroscientists cannot simply be a strange but agent-preserving link in this chain. At this point, the neuroscientists no longer seem to behaving like any sort of prosthetic at all.

accept the absurd alternative that Diagram 2a depicts an agent, it seems clear that Plum1 is not morally responsible on a literal interpretation of McKenna's prosthetic story because it fails to present Plum1 as an agent. Thus, on a strict reading, McKenna fails to achieve both his goal to present Plum1 as an agent and his goal to present an interpretation which could be used to support a hard-line reply to the 4-CA. Now, this should not overshadow the fact that the strict reading supports a viable interpretation of Case 1. However, because it is clearly one in which the neuroscientists subject to Plum1 to suppressive manipulation, the compatibilist can appeal to the same compelling soft-line reply she gave to the other viable interpretations of the argument.

By parity of reasoning, a soft-line reply is also fitting in the case where the neuroscientists faithfully offer their independent causal input to jointly cause Plum1's states. Once again, the fact that they happen to use their causal powers in a faithful way does not create the causal integration required for Plum1 to be an agent; this interpretation, too, represents the neuroscientists as subjecting Plum1 to agent-undermining, suppressive manipulation. There seems little option but to admit that our *prima facie* intuition was misleading in this case, given that an individual cannot be morally responsible unless he is as an agent. As a result, the compatibilist must abandon the hard-line reply here as well, opting instead for the response that Plum1 is not morally responsible in this case because he is not an agent. So, once again, the compatibilist can adopt the same soft-line reply given to the other viable interpretations of the 4-CA.

Upon review, then, even after the compatibilist addresses the need to reject some of her *prima facie* judgments about Plum1's moral responsibility, the Soft-line Solution to the 4-CA remains as strong as ever. The only difference is that the Soft-line Solution is now constituted by four instances of the same soft-line reply rather than two. Thus, after a grueling search for an interpretation of the 4-CA on which it poses a threat to compatibilism, we can finally conclude that there is none to be found.

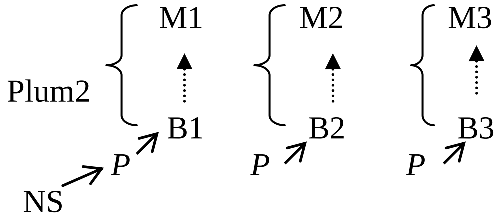
7. The 3-Case Argument

The reader might wonder why the bulk of this paper is devoted to Case 1 given that Pereboom clearly states that Case 1 is a disposable part of his argument. The reasons are simple: Case 1 is easier to work with and all of the problems in Case 1 are inherited by Case 2. So, the arguments offered above against the 4-CA are equally successful against the remaining "3-Case Argument" (3-CA) which is based upon Case 2.

Case 2 is more difficult to understand than the first case for it includes an additional feature, the so-called "programming", which is woefully under-described. Still, based on Pereboom's commentary on Case 2, it is clear that Case 1 provides the guidelines for interpreting Case 2: Case 2 simply *is* Case 1 with a time lag. Pereboom adds the time lag to Case 2 precisely because he predicts that a compatibilist might come along who has worries about Plum1's agency. Pereboom incorporates the time lag to appease such compatibilists, but is adamant that the time lag does not change anything of consequence. Reflecting on the small addition, Pereboom asks: "could a time lag between the manipulators' activity and the production of the relevant states in the agent plausibly make a difference as to whether the agent is morally responsible? (…) By my intuitions, such a time lag, all by itself, could make no difference as to whether an agent is morally responsible" (Pereboom 2002: 113). I could not agree more with Pereboom on this point, but of course therein lays the problem.

Merely adding a time lag between the neuroscientists' actions and the murder, such that "all the manipulating activity occurred during one time interval and, after an appropriate time lag, the relevant states were produced in the agent" (Pereboom 2002: 113) does not, all on its own, produce a morally relevant difference between Plum1 and Plum2. Presumably, then, since a difference in agency would be a morally relevant difference, it must be that Plum1 and Plum2 have the same status with respect to agency despite the presence of the time lag. But, given our earlier conclusion that Plum1 is not morally responsible due to his *lack* of agency, Pereboom's own reasoning suggests that Plum2 should fare no better. Indeed, as the following diagram of Case 2 represents, the problematic aspect of Case 1 which I emphasized above, the *state-by-state* control that undermines Plum1's agency, is still present in Case 2:

2b. Plum as Causally Regulated by Neuroscientists' Program (Case 2):



So, while in Case 1 there was a tension between the neuroscientists and Plum1, in Case 2 there is an exactly similar tension between Plum2 and the neuroscientists' programming. In Case 2, the program must regulate Plum2's behavior, state by state and moment by moment, throughout his life, just as the neuroscientists directly regulate the behavior of Plum1. Thus, we can see that the suppressive manipulation which prevented the victim from being a candidate for moral responsibility in Case 1 is also present in Case 2.

I suspect that those who reject my interpretation of the causal relations in Case 2 will accuse me of misunderstanding the nature of the programming that the neuroscientists have implanted in Plum2 to do their dirty work. However, while there may seem to be ample room for debate about the nature of the programming, I believe the constraints on interpreting the programming are more limiting than it may first appear. Once one has Case 1 (the acknowledged template for Case 2) clearly in mind, it seems clear that the programming given to Plum2 must be *additional* to the basic programming that must have been present in Plum1. Although we have seen that Pereboom’s stipulation that Plum1 is an agent cannot be upheld because the neuroscientists undermine Plum1’s capacity for agency through their suppressive tweaking, it still seems reasonable to imagine that Plum1 is designed in such a way that, at the very least, Plum1 *would have been an agent* had the neuroscientists simply left him alone after his creation. From the fact that the neuroscientists need to send constant radio signals in order to carry out their manipulation of Plum1, it seems clear that the basic programming that was required to make Plum1 a functioning instant agent was not sufficient to provide the neuroscientists with the control over Plum1 that they desired. This suggests that the programming discussed in Case 2 must do something more than the basic programming given to Plum1; it must be something which allows the neuroscientists to get the thoughts and behaviors that they want from Plum2 on a moment-to-moment and state-by-state basis without the hassle of constant moment-by-moment monitoring and tweaking. In other words, the programming in Case 2 is designed to carry out the same type of suppressive manipulation that was achieved by the neuroscientists in Case 1, a type of manipulation that (assuming my arguments have been successful) always undermines agency. Ultimately, Pereboom’s description of Case 2, informed by our understanding of Case 1, seems to leave little room for doubt: Plum2, like Plum1, is the victim of suppressive manipulation and is therefore not an agent. Consequently, the 3-CA offers no threat to compatibilism.

8. The *New* 3-CA and Beyond

Of course, were a proponent of the 3-CA to jettison Case 1 from consideration, then she would immediately be free to interpret Case 2 however she likes. So, even if I have provided successful criticisms of the available interpretations of Case 1 and Case 2 on a strict reading of Pereboom’s argument, there may be some alternative interpretation of Case 2 that deserves attention because it can avoid all of the foregoing criticisms of the 3-CA.

On the most (perhaps only) plausible reinterpretation of Case 2, one could take the programming that the neuroscientists give to Plum2 to be nothing additional to the minimal amount of programming that would be required for an instant agent to function just like a normal human in a deterministic world.[17] That is, one might argue that Plum2 is the nonhistorical duplicate of Plum4, i.e. a normal human in a deterministic world—not only in terms of mental and physical states as discussed earlier, but in terms of agency as well. Assuming this interpretation of the programming, there would be no grounds to conclude that Plum2 is any less an agent than Plum4, nor would there seem to be any obvious reason to think that Plum2 could not satisfy the compatibilist integrationist conditions. Presumably, the proponent of this version of the 3-CA—let us call it "The New 3-CA"—would then point out that most people do not believe that an individual created in this strange way could be morally responsible for his actions. Thus, such an interpretation seems to provide all of the necessary components for avoiding the criticisms hitherto presented in this paper while satisfying the needs of a successful manipulation argument.

I openly admit that the New 3-CA avoids my criticisms of the original 3-CA and 4-CA. However, I contend that the New 3-CA is not merely a case of making the old argument better; rather, it is a new and better argument. Currently, compatibilists have no principled or systematic way of individuating manipulation arguments, which means that they have no good way of distinguishing between a shift from one *version* of an argument to another and a shift from one *argument* to another. I believe that the best and most natural way of individuating manipulation arguments is on the basis of the specific type of manipulation they employ, as it seems that all

[17] While the term 'instant agent' is typically used to refer to cases of *ex nihilo* creation, like Swampman, I'm assuming that Frankenstein-like creations like Plum1 and Plum2 also fall uncontroversially under this heading when they are said to awaken into life with all they need (physically, epistemically, and metaphysically) to be agents.

arguments involving the same type of manipulation will be subject to the same criticisms. Now, the type of suppressive manipulation described in the original interpretations of Case 1 and Case 2 is distinctive precisely because the manipulators continue to causally infect the states of their victim, moment by moment and/or state by state. In fact, the type of manipulation Plum1 and Plum2 are subjected to—what I would call suppressive "Create & Tweak Manipulation" because of the on-going involvement by the manipulators—is not employed in any other of the well-known manipulation arguments.

On the other hand, there are already a large number of manipulation arguments which involve the creation of a so-called "instant agent" who is immediately released into the world after his creation and is not tinkered with any further by his creators—what I refer to as "Create & Release Manipulation". This type of manipulation is found, for example, in familiar cases developed by Alfred Mele ("Fred"), David Zimmerman ("Sean Young"), and Michael McKenna ("Suzie Instant").[18] In fact, part of the reason that compatibilists have been so troubled by the 4-CA is that they have been unable to extend their criticisms of the commonplace Create & Release manipulation arguments—where I happen to think hard-line responses alone are often adequate—to the Create & Tweak manipulation employed in the 4-CA. That is, it seems that the resiliency of the original 4-CA has come by way of the novel type of manipulation it employs, for it requires an equally novel response. Thus, once the defender of the 4-CA *abandons* the unique, suppressive Create & Tweak manipulation employed in the original 4-CA in favor of the non-suppressive Create & Release manipulation described in the New 3-Case Argument, he in

[18] See David Zimmerman's discussion of "Sean Young" in "Born Yesterday: Personal Autonomy for Agents without a Past" (1999); Alfred Mele's discussion of "Fred" in *Autonomous Agents* (1995), and again in *Free Will and Luck* (2006); and Michael McKenna's discussion of "Suzie Instant" in "Moral Responsibility & Globally Manipulated Agents" (2006).

fact concedes that the original 4-CA must be discarded—and, in that case, the New 3-Case Argument can be immediately relegated to the already burgeoning collection of manipulation arguments which employ Create & Release Manipulation.

Now that we have the beginning of a taxonomy of manipulation arguments, the compatibilist can confidently eliminate the entire category of suppressive manipulation arguments (such as the original 3-CA and 4-CA) from future debate. The task remaining for the compatibilists is to complete the taxonomy so that specific responses to each type of argument can be developed. Hopefully, by following this strategy the compatibilists will be able to force their opponents to retreat to an ever-smaller collection of arguments until no viable options remain.

9. Conclusion

Taken together, I believe my arguments not only show that the Four-Case Argument fails to reveal any inadequacy in contemporary compatibilism, but they also show that no future manipulation argument that employs suppressive manipulation will have any hope of succeeding. Not to be overlooked is the fact that my critique of the Four-Case Argument does not depend on the acceptance of any particular free-will machinery or any special theory of moral responsibility. In other words, the solution offered in this paper is not a mere circling-of-the-wagons defense of a particular version of compatibilism; it is designed to be a thorough-going refutation of Pereboom's argument. I grant, however, that even if my arguments are sound, there is still a great deal of work remaining for the compatibilists. Even if suppressive Create & Tweak manipulation arguments have been categorically defeated, there are many other types of manipulation arguments which are not subject to the same criticisms, and the majority of these still go without a satisfying response.

CHAPTER THREE

REDEFINING ‘DETERMINISM’

> But ‘determinism’ must, if violence is not to be done to every traditional association that word has, be used to refer to the thesis that there are no such [actual-sequence-K.D.][1] alternative possibilities.
>
> -Peter van Inwagen (1983: 86)

1. Introduction

Traditionally, the type of determinism of central interest in the free will debate has been *causal determinism,* and it has long been assumed that if such determinism is true, it is true of the entire world. These days, the assumption that causal determinism is best understood as a doctrine about the entire world remains strong, but definitions that make any appeal to the concept of causation are becoming increasingly rare. Even *The Stanford Encyclopedia of Philosophy* entry entitled “Causal Determinism” (Hoefer 2010) forwards a world-based definition of causal determinism that avoids any allusion to causal relations. Looking specifically at contemporary free will literature, the dominance of causation-free definitions of causal determinism is directly

[1] While van Inwagen does not employ the term “actual-sequence” with respect to alternative possibilities, this term provides an apt description of the type of alternatives he describes in the passage from which this quotation is taken. In the larger passage, van Inwagen describes a sequence of events which takes place in the actual world A and says that given that this sequence of events takes place in A, if the laws at A are deterministic, then nothing could have happened in A other than what did happen; there is no possible world at which the laws are the same (deterministic) laws as in A and *one* event from the actual sequence of events takes place but not *every* event in the actual sequence takes place. In other words, given the *actual* sequence of events, the truth of determinism at A entails that there is no alternative way that the world could have gone—i.e., there are no “actual-sequence” possibilities for the world.

attributable to Peter van Inwagen, and the formulations of determinism he provided in *An Essay On Free Will* have since become orthodox. Instead of causal laws, van Inwagen understands determinism in terms of the world's natural laws.[2] Referring to one of his most popular definitions, he asserts: "The reader will note that the horrible little word 'cause' does not appear in this definition. Causation is a morass in which I for one refuse to set foot. Or not unless I am pushed" (1983: 65).

Now, I have no interest in pushing for a return to *causal* determinism, and I agree that a definition of determinism given in terms of natural laws is best. However, I also think that we should stop allowing definitions which make no attempt to capture the nature of deterministic *causal* relations to masquerade as expressions of *causal* determinism. Rather, I think we should acknowledge and embrace the transition from causal determinism to what we might call "natural-law determinism" and treat these as two distinct doctrines (at least until the true nature of their relationship is made clear). When we finally look at natural-law determinism in its own right, I believe that, contrary to what van Inwagen says in the epigraph, we will see that natural law determinism allows room for actual-sequence possibilities, a fact that has long been hidden by the muddled transitional working-definitions of determinism that have dominated the literature for the last few decades. In the end, I hope to show that van Inwagen's view of the world and the orthodox view of determinism are out-dated; the time has come to rethink what it means for an event to be determined.

In the first half of this paper, I argue that the assumption that there can be only one set of natural laws in the grand history of the world—an assumption that van Inwagen and so many others seem to make—can no longer be taken for granted. As we shall see, contemporary

[2] Van Inwagen understands natural laws, in turn, as propositions which have the feature of being natural laws (1983: 60-1).

cosmologists tend to agree that the world is full of distinct universes, and some even contend that each has its own distinct set of natural laws. Drawing upon such theories, I offer a counterexample to van Inwagen's most popular formal definition of determinism, showing that van Inwagen dramatically overstates the domain in which determinism must hold. I then turn to van Inwagen's other formal definition, and provide two counterexamples to it. Along the way, I also demonstrate that, contrary to popular belief, van Inwagen's two formal definitions are not equivalent. Most significantly, though, I provide two independent arguments for the surprising conclusion that "determined" events may not occur, for it is beyond the scope of deterministic natural laws to ensure either the *existence* of one unique future or the *existence* of one unique past. This leads to a new vision of determinism, and I point to two on-going debates in the free literature that must proceed differently in its wake. Finally, I address those who wish to persist in supporting the orthodox view of determinism expressed in the epigraph despite the arguments I present. For those loyal to van Inwagen's formulations of determinism, I point out that even he acknowledges that (strictly speaking) deterministic natural laws do not rule out every actual-sequence possibility or guarantee that every event determined by such laws must actually occur.

2. Determinism's Domain (A Word About van Inwagen's World)

While clearly in the spirit of the traditional Laplacean vision of determinism, van Inwagen's most popular formal definition breaks sharply from tradition by eliminating all mention of causal relations.[3] According to van Inwagen, determinism can be understood as the conjunction of the following two theses:

[3] For those unfamiliar with Laplace's famous formulation, it goes as follows: "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes" (1820/1951: 4).

> (a) For every instant of time, there is a proposition that expresses the state of the world at that instant;
>
> (b) If *A* and *B* are any propositions that express the state of the world at some instants, then the conjunction of *A* with the laws of nature entails *B*. (1983: 65)

If this definition, which I will call the "First Formal Definition" (FFD), were correct, it would mean that there is only one possible domain in which determinism can obtain: the whole world. The assumption that determinism is best defined in terms of the world is quite common, although it is not entirely clear why. In a rare defense of why those working in the free will debate should employ definitions that assume determinism holds world-wide, Carl Hoefer (2010) seems to suggest that the only alternative is to opt for a definition in terms of individual events, but he rejects such an alternative definitions on the grounds that they would mask the features of determinism that are most relevant to the free will debate. In his concluding remarks on the matter he says: "(W)e have a number of good reasons for sticking to the formulations of determinism that arise most naturally out of physics. And this means that … we are looking at how *everything* that happens is determined by what has gone before" (2010). I completely agree with Hoefer that the definition of physical determinism we adopt for use in the free will debate should be informed by our best physics, but this is precisely why I think that he and van Inwagen have wrongly identified the proper domain for natural law determinism. When we look to our best physicists for guidance on this matter, the lesson implied by their work is clear: natural law determinism should not be understood as a doctrine about the entire world.

Of course, the term "world" is used to refer to many different things in normal English, so it is imperative that we start by clarifying the definition that is operant in FFD. From the wide range of domains from which he could choose, van Inwagen borrows a definition given by Peter Geach, and describes the world as "the upper limit of the series: the solar system, the galaxy, the system of galaxies…" (1983: 81).[4] As it happens, many eminent cosmologists now believe that this series ends in a multiverse.[5] In a multiverse-world, more than one universe exists and, at least on some theories, a distinct set of natural laws governs the goings-on within each universe. A universe's natural laws, we are told, are fixed sometime in the early stages of the emerging universe, and the natural laws which arise need not be the same in every universe of the multiverse. However wild, the type of multiverse-world described above is at least metaphysically possible. Indeed, even van Inwagen seems open to the metaphysical possibility that more than one physical universe could exist at any given possible world. In "Indexicality and Actuality", van Inwagen describes the world as "a concrete object—this huge thing that astronomers investigate, and which we find ourselves within and parts of", and later, while attempting to explain the ontological difference between the world and possible worlds, says:

> (T)here is only one cosmos (or, *even if there are many cosmoi*—many enormous closed causal systems—they are every one of them contingent objects and it should seem that there might have been just one—or none), but there are, and are necessarily, many ways things could have been" (my emphasis) (1980:406).

Once we accept that multiple universes (cosmoi) are metaphysically possible, it does not seem significantly more extravagant to posit that the natural laws are different in each. However, assuming that natural laws are sometimes restricted to a subdomain within the world, there will be sets of deterministic laws which cannot be identified as such by FFD.

In order to see the problem with FFD more clearly, let us take a look at a universe *U* which exists at the possible world *w*. In *U*, there is a set of natural laws which account for the goings-on in the universe during all times at which it exists. Let us use the term '*L*' to denote the

[4] I do not wish to be side-tracked by issues in philosophy of language here. For any reader who would prefer to preserve the term "universe" to refer to all of physical reality, I contend they can replace "universe" with their preferred term for this type of physical sub-world system without this having any affect on the arguments presented in this paper.

[5] For example: Stephen Hawking (2010), Roger Penrose (2010), Martin Rees (1997), Andrei Linde (1994).

proposition describing these laws. Let us also assume, for the sake of argument, that FFD is an adequate definition of determinism and the laws expressed by '*L*' satisfy FFD, which is to say that *L* together with *A* entails *B*. Given that FFD reflects the orthodox view of determinism, it seems reasonable to say that the laws described by '*L*' are deterministic. However, we imagine a possible world at which an exactly similar universe exists and has the same laws, and yet FFD *fails* to identify the laws of that universe as deterministic.

For instance, consider the nearby possible world *v* at which there exists a universe *U1* which is qualitatively identical to *U* and that that laws denoted by '*L'* also obtain in *U1*. Moreover, it is true at *U1* that when *A* and *B* are each propositions that express the state of *U1* at a time, the natural laws of *U1* are such that a proposition *L1* expressing those laws together with A entails *B*. Once again, we seem to be on track to reach the conclusion, based on FFD, that the laws at *U1** are deterministic. However, the physical world that exists at *w* includes more than just *U1**. In the simple multiverse at *w*, *two* universes exist in the history of the world, although only one universe exists at any given time. *U1* was born with a bang and ultimately dies in a big crunch, after which a new bang brings forth a new universe, *U2*. Now, as it happens, *U1* and *U2* are each governed by a different set of natural laws, expressed by the propositions *L1* and *L2*, respectively. At the big bang birth of *U1*, the laws described by '*L1*' emerge and account, thereafter, for the evolution of the physical world until the "death" of *U1*, at which its laws break down. Then there is another big bang and *U2* emerges. Thereafter, the world (which is now consists of just *U2*) is governed by a different set of laws, those described by *L2*.

In the scenario described above, the laws described by '*L1*' do not account for the goings-on outside the temporal-spatial boundaries of *U1*, which means that the laws described by '*L1*' are irrelevant to the chain of events which unfolds in *U2*; likewise, the laws described by '*L2*' do not hold in *U1* and, so, are irrelevant to the goings-on in *U1*. So, in a world like this one, when *A* describes a timeslice of *U1* and *B* describes a timeslice of *U2*, these propositions describe states of affairs that are not related to each other by the natural laws described by either '*L1*' or '*L2*'. Nonetheless, like *U1*, the natural laws of *U2* are such that if *A* and *B* are each propositions that express the state of *U2* at a time, then the natural laws at *U2* are such that *L1* together with A entails B. Given these descriptions of the laws at *U1* and *U2*, it hardly seems that the laws recorded by '*L1*' or '*L2*' could be *indeterministic*—yet, because the entailment posited in (b) of FFD is false of the physical world at *w*, this just what FFD implies. So, while it seems that our intuitive understanding allows us to accept that the laws described by '*L1*' and '*L2*' are deterministic, the definition of determinism proposed in FFD does not. Furthermore, FFD implies that the laws which obtain in *U* are deterministic while those in *U1* are not, even though the natural laws which obtain in *U* and *U1* are the same. Thus, it seems that FFD is in need of repair.

In order to salvage an entailment thesis that is in the spirit of (b), it seems that we must require that both *A* and *B* describe timeslices of a single universe. Thinking along these lines, we might revise van Inwagen's entailment thesis as follows:

> (b') If *A* and *B* are any propositions that express the state of the world during times at which a discrete universe *u* exists, then the conjunction of *A* with the laws of nature entail *B*.

According to the resulting version of FFD, the laws of nature of a universe *u* are deterministic if and only if theses (a) and (b') are true. It seems that (b') would allow us to identify *U*, *U1,* and *U2* as deterministic. Indeed, (b') would allow us to identify the laws of one universe as deterministic even when the natural laws of other universes in the series are not. However, (b') does not do as well when the world is a bit more complicated. If, for instance, *U1* and *U2* are parallel universes, existing (CLARIFY: “in some meaningful sense”) simultaneously, at least some propositions describing the state of the world when *U1* exists will also describe *U2* because (b’) does not demand that *A* and *B* express states of *only* one universe. This is important because if we posit that the laws denoted by ‘*L2*’ are indeterministic, then no entailment will hold between propositions describing the states of *U2*. So, when *A* and *B* express states of the world including the states of *U2*, the entailment between *A* and *B* will fail even when the laws of *U1* are deterministic. Thus, the amendments found in (b’) are insufficient to salvage FFD.

By now it should be clear that we must move away from defining determinism in terms of the world, favoring instead a definition that is given in terms of a single universe:

> (b’’) If *A* and *B* are any propositions that express the state of *just one universe in* the world, then the conjunction of *A* with the laws of nature of that universe entails *B.*

With (b’’), I believe that we have, finally, restricted determinism to the proper domain. We can now see that Hoefer was wrong to suggest that the only relevant alternative to world-wide determinism was a problematic definition in terms of individual events. By narrowing the domain to a single universe, we have identified a non-arbitrary subsystem of the world within which everything follows as a matter of natural law and outside of which the laws do not apply. Thus, by defining determinism in terms of a universe, we understand determinism in terms of the relevant domain for discussions of free will, i.e. the largest domain in which everything that happens is determined by what has gone before as the result of natural law.[6] Since the resulting definition allows us to focus on all and only those events governed by the natural laws of a

[6] Of course, it would also be possible to define determinism in terms of an arbitrarily small closed system within the universe: Einstein, for example, defined determinism in terms of such systems (see Byrne 1981: 914).

System-based versions of van Inwagen’s definitions can also be found in the free will literature (eg., Nahmias, Coates, Kvaran 2007: 215). However, it seems likely that definitions in terms of arbitrary systems would often fail to capture all of the relevant law-based relations between events in or states of the universe and, so, as Hoefer points out, such definitions would not be suitable for the free will debate.

universe, it should assuage Hoefer's worry that opting for a domain smaller than the world would obscure some philosophically interesting deterministic connections between events.

At this point, I expect that some readers will question whether any significant advance in our understanding of determinism has been made. To these readers, my critique may seem to boil down to the minor complaint that there has been a shift in the meaning of 'universe' and, so, van Inwagen's use of the terms 'world' and 'universe' as synonyms is outdated, and that a good definition must be given in terms of the latter. Moreover, one might argue that the free will literature is already speckled with van Inwagen-style definitions employing entailment theses like (b''), given that others—Alfred Mele (e.g. 2010), Eddy Nahmias (e.g. 2011), and Adina Roskies (e.g. 2006), to name a few—routinely opt to use the term 'universe' rather than 'world' in their van Inwagen-style definitions of determinism. Given the scientific bent to their research, it might be thought that they do so out of recognition that the world might include distinct universes with their own distinct laws.

True, several leading philosophers have taken the liberty of altering FFD to create their own similar van Inwagen-style definitions, but I would like to point out that to my knowledge none of these other van Inwagen-style definitions are accompanied by an argument to show that the shift from 'world' to 'universe' is a philosophically relevant change.[7] At best, I believe that the term 'universe' is preferred because it makes it easier to refer to the world without generating confusion between *the world* and *the actual world*. More importantly, though, my discussion of why world-based definitions of determinism are false has been designed to do more than defend

[7] Mele typically seems content to use the term 'universe' to refer to the entire world. However, in at least one place he notes that "Some readers will wish to insert 'after the Big Bang' between 'instant' and 'exactly'" into one of van Inwagen's definitions of determinism ("The thesis that there is at any instant exactly one physically possible future") (Mele: 142). Notably, though, while referring to the Big Bang shows that Mele is sensitive to the idea that natural laws do not hold *at all times* in the history of the universe, said reference does not show that Mele has made the more radical break from tradition that I recommend, i.e. thinking of the universe as something less than the entire world.

yet another van Inwagen-style definition; in fact, it grounds a more substantial and surprising critique of the orthodox view of determinism.

3. Mayhem in the Multiverse

Now that we are more comfortable thinking about the possibility of one deterministic universe existing among other universes, another intriguing possibility immediately presents itself: What would happen if the universes were to *collide*? It seems that the possibility of collisions between discrete universes is quite widely accepted among cosmologists, although there is disagreement about what would happen to the universes as a result. (As if the mere possibility of collision were not interesting enough, Stephen Feeney and his research team claim to have found *evidence* that our very own universe has survived a collision with another universe (Feeney et. al., 2010).) Of course, whether or not these scientists are right about our world is not important in the present context; what matters is that their work suggests that collisions among universes in a multiverse are at least metaphysically possible. This fact has surprising implications for our understanding of determinism.

As mentioned earlier, if natural laws govern only within the boundaries of an individual universe, then natural laws are irrelevant to events that take place outside that universe. This means that, regardless of whether the laws are deterministic or not, the natural laws of a given universe do not govern the relations it has with other universes in the world. As such, it seems that the natural laws of a universe do not prevent the universe from colliding with other universes. If so, the entailment posited in (b’’) and all definitions of determinism employing something like it—e.g. those promoted by Mele, Nahmias, and Roskies—are in trouble.

Consider a scenario in which the parallel universes *U1*and *U2* described above do not remain parallel. Imagine instead that *U1* collides with *U2*, leaving what is colloquially known as a “cosmic bruise” on *U1* as a result of the interaction. Such bruises (a concentric wave pattern in the cosmic microwave background of universe) are states of a universe that are caused jointly by prior events in that universe together with its natural laws *and events that occur outside the boundaries of the universe*. To recognize the possibility that a universe can undergo changes as the result of colliding with another universe, then, is to recognize that the state of a universe is not always a mere function of the natural laws and the past facts of the universe. So, while the entailment in (b’’) holds when a universe with deterministic laws is allowed to evolve without disruption, it is beyond the scope of those laws to *ensure* that a universe will always unfold without interference. For quite similar reasons (due to certain complications arising from his theories of relativity), Einstein, in his own Laplacean definition of determinism, appealed to “isolated” systems to block off all potential external influences that would threaten to disrupt the deterministic evolution of a system. That is, Einstein adds this as an idealizing stipulation, not because it is a feature of deterministic systems that they *must* continue to evolve without external interference.[8]

One might wish to follow suit and update (b’’) by asserting that determinism holds in an *isolated* universe, but I believe that this would be a mistake. First of all, adequately defining “isolated” would be an extremely difficult task. Second of all, I believe that employing the term “isolated” would promote the misunderstanding that the isolation of a deterministic universe is somehow guaranteed by the laws. Since adding the term “isolated” to a definition of determinism

[8] For an interesting discussion of the problems with Einstein’s definition given his theories of relativity, see Byrne 1981. Regarding the isolation of systems, Byrne makes the following comments which I take to support my position here: “A … possible way of realizing the isolated system would be to construct some sort of ‘container’ to screen off all external influences. However, if we stipulate that no container of infinite potential could be constructed in reality, it is necessary to admit the possibility of *some* external physical influence which could breach the ‘container’. … While it is true that for any given ‘container’ it is possible to calculate with great precision what sorts of effects could break it, one could not predict whether or not such an effect is in a position to do so prior to time $t = t_0$” (1981: 926).

is, ultimately, adding a *ceteris paribus* clause to the definition, perhaps we should make use of a straightforward *ceteris paribus* clause to express that deterministic relations hold only in the absence of any funny-business from beyond the boundaries of the universe. The resulting definition would look something like the following:

> (a) For every instant of time, there is a proposition that expresses the state of the world at that instant;
>
> (b*) If *A* and *B* are any propositions that express the state of a single universe at some instants, then, *ceteris paribus*, the conjunction of *A* with the laws of nature entails *B.*

I believe that the addition of a *ceteris paribus* clause to FFD would serve as a valuable reminder that the laws of nature can be "trumped". I also predict that such a *ceteris paribus* clause would be the source of much confusion. So, while the above definition, which I will refer to as "FFD_{CP}" does have its appeal, we can do better.

I will return to the project of amending FFD in Section 5 below, after a brief discussion of van Inwagen's second formal definition of determinism. Before moving on, though, I would like to pause in order to point out that the above scenarios show more than the fact that there is need for some adjustment to FFD. The truly surprising lesson from the multiverse scenarios I have discussed is that *deterministic natural laws allow that a determined event need not occur*. Looking back to the parallel universe scenario, we see a possible world at which the full history of *U1* is allowed to unfold in accordance with its natural laws, i.e. the future that was *determined* to happen according to the laws and the initial state of *U1*. As such, this scenario gives us a clear vision of the future that would have unfolded in the bruise-suffering *U1* if it had not collided with another universe. Comparing the two possible histories of *U1,* then, we see that events which are determined to take place in a universe might, nonetheless, not occur.

4. The World and the Actual World

In addition to FFD, van Inwagen offers another formal definition of determinism which I will call the "Second Formal Definition" (SFD):

> S*xy*: *x* shares a slice with *y*;
> N*xy*: *x* is nomologically congruent with *y*;
>
> We shall also employ a one-place predicate, 'D':
> D$x=_{\mathrm{df}}$ (∃*y*)(S*xy*) & (*y*) (S*yx* & N*yx*. ⊃ *y*=*x*)
> 'D*x*' is read, '*x* is deterministic'. (1983: 83)

While FFD defined determinism in terms of the world, SFD defines determinism in terms of possible worlds. Van Inwagen summarizes this definition by saying that "a world is deterministic if that world itself is the *only* world that both shares a slice with it and has the same laws of nature it does" (1983: 86). According to van Inwagen, FFD and SFD are alternative expressions of the same thesis—and the logically equivalence of FFD and SFD is widely accepted.[9]

This shift from *the physical world* to *the actual world*, is often taken lightly, but there is a world of difference between the two. Van Inwagen takes care to point out that the former is the physical thing in which we live and breathe, while the latter is an abstract object, ontologically indistinct from other possible worlds:

> Since possible worlds are possibilities and possibilities are abstract objects, possible worlds, including the actual world, are abstract objects. Therefore, what philosophers call "the world" (…) is not the same object as the actual world. The world is the universe, or the cosmos, or what Professor Geach has called "the upper limit of the series: the solar system, the galaxy, the system of galaxies…" (1983: 81).

Notably, van Inwagen does not provide a citation for this quotation, but it seems to be taken from a short essay on Aquinas (Geach 1961: 111). The context of the quotation is interesting in the present discussion because Geach forwards this definition in a discussion of Aquinas's view of

[9] For instance, FFD and SFD are presented as logically equivalent expressions of determinism by John Martin Fischer and Mark Ravizza (1998:14) and Kadri Vihvelin (2011; 2008) .

the relation of God to the physical world, claiming that Aquinas believed that the world (*mundus*) is a ‘great big object’ that was made by God, but God is not himself part of the world. Taken out of context, Geach’s definition of the world is non-committal about the relation between God and the world, but it seems that van Inwagen employs Geach’s definition precisely because of its connection to Aquinas’s views on this matter. Much earlier in *An Essay on Free Will*, van Inwagen describes Nature as the “enormous object that the natural sciences investigate”, and says that he believes that there could be an agent who is “superior to and is not a part of Nature” (1983: 14). Thus, there is considerable evidence that van Inwagen believes that interesting things can exist beyond the boundaries of the physical world.

Moving forward under the assumption that van Inwagen is right, and that a supernatural agent could exist outside the physical world, it becomes easy to illuminate one obvious problem with the formal statement of SFD. All we need do is imagine two possible worlds at which there is a supernatural creator, call him “Creator”. In each of the two possible worlds, *W1** and *W2**, Creator exists and makes a physical world and the evolution of the world is governed by a set of natural laws (to which Creator, as a supernatural being, is not himself subject). The physical worlds that exist at *W1** and *W2** are qualitatively identical; nonetheless, there is a small difference between the two possible worlds: a single thought had by Creator at *W1** is not had by Creator at *W2**. According to SFD, this small difference in the states of Creator at these two possible worlds is sufficient to establish that the natural laws which obtain in the physical universes at *W1** and *W2** are not the same. In my view, the fact that SFD has this consequence suggests that SFD is wrong.

The above critique targets SFD *as stated*, but van Inwagen later indicates that his formal statement of SFD is not complete. Specifically, van Inwagen says that even though SFD defines determinism in terms of possible worlds, he *intends* for the focus to be on the physical world that exists at each possible world: "When I talk of the state that a possible world *w* is in at time *t*, I am to be taken as talking about the state that, at *w*, *the* world—the cosmos, the universe—is in at *t*" (emphasis in original) (1983: 84). And, later he explains that the relevant timeslices are the timeslices of the physical world alone, and not the slice which include the grander collection of things that exist at any given possible world: "(I)f we are willing to think of a possible world (*strictly speaking, to think of the universe that exists in that world*) as a compact sequence of instantaneous three-dimensional 'slices', then we may say that the indistinguishability relation holds between two worlds just in the case that they have a slice in common" (my emphasis) (1983: 85). Admittedly, once we integrate van Inwagen's passing comment about how SFD should be understood 'strictly,' SFD will no longer be subject to the attack I have launched against it.

Even though my preliminary critique of SFD fails, I believe it is worth noting that van Inwagen's formal statement of SFD cannot be taken at face value, as I believe this fact is often overlooked. Say, though, that we undertake the project of adding to SFD, in all the right places, the phrase 'the physical world.' We would, thereby, develop a more complete formal expression of determinism along the lines of SFD, call it 'Strict-SFD', that better reflects the conception of determinism that van Inwagen had in mind.

By adopting a slight variation on the Creator story that I employed above, I believe it is possible to show that FFD and Strict-SFD are not equivalent either. Consider, for example, the possible world *W* described by Joseph Keim Campbell at which the first state of the physical world is a complex state but the world has no creator:

> Suppose that *W* is a determined world such that some adult person exists at every instant. Thus, *W* has no remote past. At its first moment of existence lived Adam,

> an adult person with all the knowledge, powers, and abilities necessary for moral responsibility. Shortly after Adam comes Eve, and the rest is history. (2007: 5)[10]

I agree with Campbell that this "Instant-Adam" world is metaphysically possible—which is to say that I have no commitments which imply that it is *impossible*—and I do not think that Strict-SFD implies otherwise. That said, Campbell's description of *W* is quite generic, so there are actually a number of possible worlds which satisfy his description. Let us start by considering the possible world *W1*. Let us assume that at *W1*, Adam's universe is governed by a set of natural laws *L1*. Let us also say that his universe suffers no external interference of any kind during its history. Since Adam must be in some determinate state during his first instant of life, let us say that he comes into being with his eyes closed, and then he opens them in the next instant.[11] Finally, let us assume that the entailment thesis of FFD is true at *W1*, so the laws described by '*L1*' are deterministic. These assumptions about Adam's first state are arbitrary; I could, just as plausibly, assumed a different story about Adam's first states. In that spirit, let us consider a nearby possible world, *W2*. At *W2*, FFD is satisfied, which means that the laws of this universe, recorded by '*L2*', are also deterministic. However, Adam comes into existence with his eyes open in the first moment at *W2*. Indeed, Adam's eyes are not just open, but open in exactly the same way as his counterpart's eyes in the *second* moment of *W1*. In fact, if *P1* is a proposition that expresses the first state of *W1* and *P2* is a proposition that correctly expresses the *second* state of *W1*, *P2* also accurately describes the *first* state of *W2*. So, as it happens, *W1* and *W2* share every timeslice except for the one described by *P1*.

[10] For those who, for issues related to the principle of sufficient reason, find the case to be more plausible when some type of creator exists, I encourage these readers to adjust the following discussion accordingly.

[11] The opening of Adam's eyes is a process that would take a few instants, of course, but I believe we can safely ignore that fact for the sake of simplicity.

According to Strict-SFD, the difference in this single timeslice means that the same set of laws cannot govern Adam's universe in both *W1* and *W2*, but I see no reason (apart from Strict-SFD) to deny that *L1* and *L2* are instantiations of the same set of deterministic natural laws. The fact that the laws hold for one fewer instant at *W2* than at *W1* does not suggest that the laws themselves are different; meanwhile, I take the fact that the laws are such that they bring about exactly the same states moving forward from the first shared timeslice as solid evidence that the laws are the same. Assuming, then, that *W1* and *W2* differ only with respect to one timeslice even though *L1* and *L2* are the same, I find that these possible worlds stand as a counterexample to Strict-SFD.

As indicated earlier, van Inwagen says that he is "not troubled" by the fact that determinism, as construed in FFD, entails a unique past in addition to a unique future—and we have seen that this is not troubling, so long as we do not understand "determined" as "determined to exist". However, van Inwagen worries that others might be troubled by laws which are future-to-past deterministic and, so, suggests that a "later than" clause be added to his definition by those who prefer a one-way, past-to-future definition of determinism (1983:65). A one-way version of Strict-SFD would not be quite as elegant at the original, but it certainly could be developed. Such an adjustment to Strict-SFD might seem desirable in the light of *W1* and *W2*, for, despite the differences in their pasts, the universes at these worlds have exactly similar futures from their first shared timeslice onward as a result of their deterministic laws. Recall, though, that it was *stipulated* that *W1* and *W2* are worlds at which the natural laws *and only the natural laws* account for the evolution of Adam's universe. What happens when we look for versions of *W* when we are not restricted by this stipulation?

When released from this stipulation, it no longer seems that every universe which shares a timeslice and natural laws of Adam's universe will have the same future onward from the first shared timeslice. Among the reasons that this is so is the fact that, as I argued in the earlier critique of FFD, the world might include more than one physical universe and these universes might interact. For instance, returning to possible worlds at which Adam exists, we might imagine a possible world *W3* at which Adam's universe suffers a cosmic bruise and, therefore, fails to share the full future that unfolds for Adam's universe at *W1* or *W2*. For those who are unconvinced that cosmic bruises are metaphysically possible, we might instead posit the result of two universes colliding is their mutual destruction or annihilation. Or, if, in general, the notion of colliding universes fails to entice, we might imagine instead a possible world *W4* at which Adam's universe was created by a very powerful being which soon tires of its creation and, so, destroys poor Adam and his universe long before the complete future determined by the natural laws takes place. We might imagine that the destruction takes the form of genuine annihilation, leaving literally nothing of the universe behind, but it would serve our purposes to imagine that this supernatural entity simply crushes the system back into a tight ball of matter (a singularity). Notably, we need not even assume that this being is a non-physical entity which exists outside the physical world; so long as it exists outside the boundaries of *Adam's* universe, it seems that the natural laws which hold within the universe would not preclude such a being from bringing its creation to a premature end.[12] So, whether our focus is on the determined past or the

[12] While completing the final draft of this paper, I was directed (by Joseph Keim Campbell, with my gratitude) to Scott Sehon's recent critique of the definition of determinism employed by van Inwagen in the Consequence Argument (Sehon 2010). Sehon argues that this definition is flawed because it implies that the existence of an "intervenionist God", i.e. a God which can intervene in the on-goings of the natural world, is logically impossible. I am sympathetic to Sehon's conclusion and find that his (comparatively narrow) critique lends support to my own. As a side note, I do not think there is much reason to hope, as Sehon does, that the Consequence Argument will no longer be sound when run with a definition of determinism (like DEV) which allows for an interventionist God and other interference with the determined causal chain.

determined future, we can see that its existence rests upon more than the natural laws and current timeslice—regardless of whether the timeslice is of a possible world, a physical world, or a physical universe—and, thus that Strict-SFD is an inadequate expression of natural law determinism.

In addition to revealing that SFD and Strict-SFD are false, the above discussion also highlights the fact that neither SFD nor Strict-SFD is logically equivalent to FFD. Looking back at FFD, the entailment in (b) describes a relation that holds between all states of the physical world. Let us assume that *W2* is the actual world, which means that the real world does not include the state described by *P1*, which means that *P1* is false at the actual world. However, because FFD demands that *A* and *B* each must describe a state of the world, the fact that *P1* does not describe a state of the world (at the actual world) means that *P1* is not a candidate for *A* or *B*. That is, since *P1* is a proposition describing a state of affairs that is not realized in the world at *W2*, the fact that *P1* is false at *W2* does not indicate that the entailment posited in thesis (b) of FFD is false at *W2*. This shows that FFD allows for the existence of physical worlds at distinct possible worlds which have the same deterministic laws but do not share every timeslice while SFD and Strict-SFD do not. This, in turn, makes it clear that these formal definitions are not equivalent expressions of the same doctrine—and without appeal to anything so controversial as the possible existence of strange things like multiverse worlds with colliding universes. As the non-equivalence of these definitions will presumably come as a surprise to many, I offer this as an independent challenge to the status quo.

5. Determinism: A Working Man's Definition

Having seen that FFD and (Strict-)SFD are not equivalent, and that (Strict-)SFD suffers from serious problems which do not face FFD, I will proceed with the project of shoring up FFD into an adequate working definition for use in the free will debate. In Section 3, I discussed the possibility of amending FFD with the addition of a *ceteris paribus* clause, but indicated that I believe a better option might be available.

I find that FFD_{CP} successfully draws attention to the often overlooked fact that natural laws can be “trumped”, as with miracles, or momentarily suspended for no reason at all. But it does so almost too well. That is, FFD_{CP} fails to emphasize that FFD offers a fundamentally correct vision of the regularity imposed on nature by deterministic natural laws, and instead emphasizes instead that, technically, just about anything can happen even in a universe with deterministic laws. In the context of the free will debate, the question most commonly asked in relation to deterministic natural laws is whether the obtaining of such laws would make it impossible for an agent (like a human being) who is part of the natural world and subject to its laws, to be free and responsible for her actions. Since, according to FFD_{CP}, anything can happen in a deterministic universe, the real tension between free will and moral responsibility one hand and determinism on the other is hidden somewhere in the depths of the *ceteris paribus* clause.

What we need, in the context of the free will debate, is a definition that shows the tension between free choice and determinism on its face. Thus, I suggest something like the following thesis to replace FFD:

> *The Thesis of Deterministic Evolution* (DEV):
>
> A universe *u* evolves according to deterministic natural laws during some interval of time *I* (where *I* is an interval including times T_0-T_{0+n}) if and only if
>
> (a) the natural laws and *only* the natural laws account for the state-to-state evolution of *u* during *I*,
> (b) for every instant in *I*, there is a proposition that expresses the state of the universe at that instant,
> (c) when *A* and *B* are any propositions that express the state of a single universe at some instants in *I*, the conjunction of *A* and a proposition L expressing the laws of nature entails *B*.

Although DEV is more complicated than its predecessor, it is still simple enough to be a working definition. DEV allows for strange possibilities like miracles, time-indexed natural laws, suspensions of natural laws, collisions between universes, etc., but without letting these strange possibilities overshadow the regularity promised by deterministic natural laws when they are, we might say, "in full effect". Third, DEV makes it easy to home in on intervals during which no such strange disruptions to the laws occur and, so, consider the implications the laws for agents who are subject to those laws without having these implications clouded by the implications of miracles or suspensions.

Put another way, some states of affairs in the universe take place as a direct *result* of the laws while others take place *in spite* of them, and I contend that a good working definition of determinism will help us separate the latter from the former. In doing so, the definition would allow the traditional debate over the compatibility of free will and determinism (whether an agent who performs some action as the *result* of deterministic natural laws can do so freely) to continue without significant interruption while opening up the door to other interesting questions that have been ignored. For instance, one might ask: If one *result* of deterministic laws is that no agents have access to the type of alternative possibilities required for free will, might there be some events that happen *in spite of the* laws—miracles or a suspension of the laws at just the right moment in an agent's decision-making, say—which preempts that result and, so, allows a person to be free? I think a good definition of determinism will allow us to ask such questions and I believe that DEV fits that bill.

6. Two Applications: Prediction and Prepunishment

On the view of determinism I have defended, it is possible for an event to be determined by the natural laws and, nonetheless, fail to take place. Put another way, I am advocating for a definition of determinism according to which there are actual-sequence alternative possibilities even in a deterministic universe. That said, I must also point out that I am not committing myself to the view that, using van Inwagen's terminology, a person has *access* to any particular (set of) non-possible worlds, i.e. that a person has the ability to get some possible world at which the laws are not in full effect to be the actual world (1983: 90). So, although robust actual-sequence alternative possibilities are not ruled out by deterministic laws, as has traditionally been assumed, it is unlikely that this new view of determinism will be of use in providing novel responses to arguments like van Inwagen's Consequence Argument.[13]

However, there are other discussions where the actual-sequence alternative possibilities left open by determinism could be important to the outcome of the debate. For instance, there is a large literature devoted to discussing the connection between determinism and predictability, and this new way of thinking about determinism has clear implications for this debate. If the facts of the past together with deterministic laws of nature do not guarantee future facts, but guarantee only what future will occur on the condition that nothing disrupts the system governed by these laws, then perfect knowledge of past facts of one universe and its laws of nature will not be an adequate basis for perfect predictions of future events in that universe. This will certainly disrupt some conclusions about the relation between prediction and determinism. As a case in point, Stefan Rummens and Stefaan Cuypers (2010) argue that there is an important distinction between agents who aspire to make predictions while embedded within the deterministic system about which he is making predictions and those who are outside the system, concluding that only the latter are able to use the function provided by the deterministic laws to predict the future. However, if I am right, even an agent external to the system would need to be armed with more

[13] Indeed, I argue that FFD is an adequate definition in the context of the Consequence Argument in "Consequences of Determinism and the Consequence Argument: A Reply to Sehon" (unpublished manuscript).

than knowledge of the natural laws and states of the universe in order to make accurate predictions about the future.[14]

Assuming that determinism does not allow for straightforward prediction, arguments which rest on the assumption that determinism supports prediction are also in trouble. For instance, consider the pesky problem of "prepunishment" that has been promoted by Saul Smilansky. According to Smilansky (2007), compatibilists have no principled way to reject prepunishment (the seemingly immoral practice of punishing an agent before he or she actually commits the crime for which the punishment is meted out). However, laying the foundation of his argument, Smilansky asks that reader to assume determinism and "complete predictability". That is, he rests his argument on the assumption that "if people's actions are determined, and we have perfect epistemic capacities, we can know ahead who will commit a crime" (2007: 347). But we can see that the assumption of complete predictability is not justified and, so, the compatibilist need not accept it. Furthermore, in the light of DEV, it seems that the compatibilist can even appeal to the same general principle behind the commonsense, libertarian principle that Smilansky suggests as a way to rule out prepunishment. That is, assuming that an agent deserves every last chance to avoid becoming a criminal (otherwise prepunishment would be involve the punishment of an innocent person, and even a compatibilist has the tools to say that this is wrong), we must wait for an agent to commit a crime *even if the natural laws of his universe are deterministic.* This is because there are any number of ways that the expected evolution of a

[14] In addition to the laws which govern the evolution of events within a given universe within the multiverse, it seems likely that there would also be some "meta-laws" which would govern, for example, how distinct universes related to each other. Assuming that are such laws, the standard view is that these laws would have to be either deterministic or indeterministic. So, if God ensured that no other beings than he would perform miracles relative to the deterministic laws of a given universe *u* and the meta-laws of the world were also deterministic, it does seem that God would be in a position to predict, based on the laws of the world, what the future of *u* would be like. Short of God, though, I do not see how any being could use the laws to make infallible predictions about what will occur within the boundaries of a deterministic universe.

universe could be disrupted—from the rippling effects of a cosmic bruise to outright destruction of the entire universe prior to the agent's commission of the crime. So, even if determinism does not allow the agent to *choose* otherwise, there is always at least one morally relevant actual-sequence alternative to the agent's committing the crime: the agent's death. When punishment is carried out on someone who is *determined* to perform a crime but never actually does so, that punishment is *uncontroversially* the punishment of an innocent person—and, again, even Smilansky recognizes that a compatibilist can reject this type of immoral treatment. Whether Smilansky's argument can be revised to handle this type of response is beyond the scope of this paper, but it certainly makes the matter more complicated and at least seems to open the door for a new type of compatibilist reply. Whether or not any of these suggestions will bear fruit for the compatibilist remains to be seen, but they do seem to be worth exploring.

7. Making Room for Miracles

In this paper, I have made several breaks from the orthodox view of determinism employed in the contemporary free will debate. In this, the final section, I would like to offer some preliminary replies to some of the lingering worries about DEV that I anticipate will be shared by many readers.

First, while I drawn upon the work of leading physicists, I contend that I am not succumbing to the "hegemony of physics" about which Ted Honderich rightly complains (2002: 462-63). My critiques of FFD and SFD do not depend on any particular interpretation or truth of any empirical theory. I contend that their authors' background commitments to controversial ontological and other metaphysical views are not relevant here. I merely assume that, put to the task, one could flesh out the details of the multiverse scenarios I have provided *in some metaphysically coherent way*. I suppose that many will find even this to be a controversial assumption. I invite the arguments which establish that there is a problem with every one of the variegated scenarios that I have forwarded in support of my claim that determined events need occur.

I would also like to emphasize that this paper is devoted entirely to shoring up the *working* definition of determinism for use in the free will debate. The definition I have proposed for this purpose, DEV, is not intended to be a definitive statement of determinism any more than van Inwagen presented FFD or SFD to fill that role. That is, as van Inwagen openly stated, he introduced FFD and SFD to express his working-definitions of determinism (1983:11); it is only because so many others have found these working-definitions to be adequate (and, presumably, because of their employment in his influential Consequence Argument) that FFD (and, less so, SFD) has become the orthodox working definition of determinism in contemporary free will literature. Also, like FFD and SFD, DEV is able to accommodate a variety of realist views of natural laws, including both governing-law theories and theories which posit that laws reduce to brute dispositions.[15] Seen this way, the seemingly disruptive implications of my thesis are tempered by the fact that my goal in this paper is really quite a modest one.

Still, I recognize that the modifications to the orthodox definition I have suggested here will be so shocking to some that they will be loath to accept that DEV is even an expression of the concept of determinism. I suspect that some will say that it is tautological that determinism rules out actual-sequence possibilities, such that accepting a definition of determinism which

[15] Again, I am working under the assumption that it is no longer feasible to assume that natural laws hold ubiquitously throughout the world because any theory of natural law which rests on this assumption would be subject to the same counterexamples I have presented to upset FFD and SFD.

Also, while it would be beyond the scope of the paper to discuss, I would like to note that (at least some) law necessitarians will reject the metaphysical possibility of some of the scenarios I have described in this paper. However, I contend that my critique of SFD shows that one can reach the conclusion that deterministic laws do not ensure the existence of a unique past or future even on the assumption that the natural laws are the same at every possible world. I would also add that even Alexander Bird, a prominent necessitarian, admits “The received and intuitive view of laws is that they are contingent” (2004: 256).

allows for such alternatives is tantamount to accepting a definition of bachelorhood which allows for some bachelors to be married men.[16] Indeed, I recognize that there are various places where philosophers make assertions about what is required of an "adequate account" of determinism (cf. eg. Fischer and Ravizza 1998: 14), and such claims, if taken literally, would rule out DEV as a candidate expression of determinism. Indeed, one might point to the epigraph of this paper as evidence that something has gone terribly wrong if we have arrived at a definition of determinism which allows for actual-sequence alternative possibilities.

However, there is evidence that not even van Inwagen would dismiss DEV out of hand, despite his comments in the epigraph. In the introduction of *An Essay on Free Will*, van Inwagen briefly discusses the possibility of miracles, a discussion that I think bears repeating:

> Now I am not one of those philosophers who think that miracles are conceptually impossible. It seems to me that if God created *ex nihilo* a spinning object, then the proposition we call 'the law of the conservation of angular momentum' would be false. Yet, it seems to me, it might be a law of nature for all that. I think I understand the notion of a supernatural being, that is, the notion of an agent who is superior to and not a part of Nature (this enormous physical object that the natural sciences investigate), and I think that the falsity of a proposition counts against its being a law of nature if and only if that falsity is due entirely to the mutual operations of natural things, and not if it is due to the action of such an "external" agent upon Nature. But it does not follow from this perhaps rather quaint thesis about the concept of *miracle* that *we* can perform miracles" (emphasis in original) (1983: 14-15).

This quote reveals that even if there are those who do not endorse the existence of supernatural beings and, so, dismiss my appeals to God's creating and destroying the world, it seems that I have at least succeeded in showing that van Inwagen would likely accept that my God-involving scenarios are possible. The more important lesson to glean from this quote, though, comes from van Inwagen's claim that "the falsity of a proposition counts against its being a law of nature if

[16] My thanks to Brad Monton for raising this worry.

and only if that falsity is due entirely to the mutual operations of natural things, and not if it is due to the action of such an 'external' agent upon Nature". This line shows that van Inwagen *accepts* that actual-sequence alternative possibilities for the world even if its evolution is governed by deterministic laws; in the present context, it is only a sidebar that he thinks that only a supernatural *agent* like God could actualize any of them.

In developing DEV, I have simply appealed to much the same principle that van Inwagen seems to have had in mind in the above quotation. That is, just as van Inwagen posits that God's influence in the world is a data point that must be handled by an adequate theory of natural laws, I think that the influence of external non-agents and other law-disrupting occurrences must be allowed for as well. I contend that DEV simply makes my commitment to this principle explicit. So, it seems that even van Inwagen, arguably the most famous proponent of the orthodox working definition of determinism in the free will debate, would not be opposed to a more careful definition in the general vicinity of DEV, i.e. a definition of determinism which allows that a determined future need not happen and that a determined past might never have been.

8. Conclusion

I began this paper by identifying the subversive shift from causal determinism to what I call natural law determinism. In the remainder of the paper, I exploited this shift to show that the orthodox view of determinism in the free will literature mischaracterizes this doctrine. I started my attack with a critique of van Inwagen's most popular formal definition of determinism, FFD, and then turned to van Inwagen's second formal definition of determinism. Not only did I show that each is inadequate as it stands, I also concluded that SFD and FFD are not equivalent. Perhaps most significantly, I argued that the truth of natural law determinism in a universe guarantees neither the existence of the unique past nor the existence of the unique future that is consistent with those laws and a given state of that universe. I reached this same surprising result from two distinct critiques, one of FFD and one of SFD. As such, my critiques of FFD and SFD provide independent grounds for the view that determined events need not occur. While a complete discussion of the many implications of this result is beyond the scope of the paper, I suggested several ways that a shift to the revised definition of determinism I defend, DEV, might be relevant to two popular discussions about freedom and determinism. I closed with a defense of DEV that revealed that even van Inwagen accepts that it is not true, *strictly speaking*, that the unique past and unique future that would follow as a function of the laws and the facts of the past must be the future that does come to pass. Taken together, I believe that the arguments in this paper provide a substantive critique of the orthodox view of determinism and a bold step towards a new and improved working definition of determinism for use in the free will debate.

CHAPTER FOUR

BEYOND THE "THREE-FOLD CLASSIFICATION"

1. Introduction

Kadri Vihvelin defends what she calls a "Three-fold Classification" of free will compatibilism, incompatibilism, and impossibilism (2011; 2008). The central purpose of this Three-fold Classification is to provide a correct characterization of the logical relationship between incompatibilism and impossibilism, a relationship which is commonly misunderstood. In this essay, I argue that the Three-fold Classification must be rejected because it provides an impoverished view of compatibilism, an untenable characterization of incompatibilism, and, so, a misrepresentation of the logical relationship between incompatibilism and impossibilism.

I begin this essay with a brief summary of Vihvelin's Three-fold Classification. I then present a novel counterexample to her preferred definition of 'incompatibilism'. Next, I provide a rough sketch of an alternative mapping of the logical landscape which better reflects our intuitive understanding of incompatibilism, compatibilism, impossibilism, and the logical relationships between these views. As part of this positive project, I also demonstrate the inadequacy of Vihvelin's preferred characterization of compatibilism—a result that has wider implications for the free-will debate, as Vihvelin endorses one of the most popular definitions of 'compatibilism' in current free-will literature. I close by offering some insight on what a proper characterization of compatibilism might look like.

2. Vihvelin's Three-fold Classification

At the center of Vihvelin's Three-fold Classification of compatibilism, incompatibilism and impossibilism is her preferred version of the free-will thesis:

> *Vihvelin's Free-Will Thesis* (VFT): (A)t least one (non-godlike) creature has free will (2011; 2008: 304).

More formally, we can represent VFT as follows:

> 'Hx' represents *x is a human-like being*
> 'Ay' represents *y is an action*[1]
> 'Fxy' represents *x freely performs y*
>
> (VFT)$=_{df}$ $\exists x \exists y(Hx \ \& \ Ay \ \& \ Fxy)$

Let us name the view that VFT is strongly metaphysically possibly true "Possibilism"; where '◊' represents strong metaphysical possibility, Possibilism is the view that ◊VFT is true.[2] So understood, Possibilism is the contradictory of what Vihvelin calls "Impossibilism", i.e., the view that it is (strongly) metaphysically impossible for free (non-godlike) agents to exist (2008: 303).[3]

"Compatibilism", Vihvelin says, "is the claim that [strongly metaphysically] possibly, determinism and the free will thesis are both true" (2011; 2008: 305).[4] As such, compatibilism, as Vihvelin understands it, implies Possibilism. Vihvelin contends that her characterization of

[1] I am using the term "action" to denote the type of activity which the average person would call an action (as opposed to a mere reflex, etc.), i.e. the type of activity that at least seems to be a *contender* for an action, and, so, for an action which might be free.

[2] In this essay, all possibility and necessity claims should be understood as making a claim about *strong* metaphysical possibility/necessity.

[3] In the wider context of the free-will debate, we will want to distinguish between the debate over whether ◊VFT is true from the debate over whether *any* metaphysically possible beings act freely (including God and other god-like beings). In this essay, I will use the term 'Impossibilism' consistently to refer to the qualified impossibilist view that it is (strongly) metaphysically impossible for a *human-like being* to be free, although I will discuss arguments which support an unqualified version of impossibilism. This point is discussed in greater detail in fn. 10.

[4] Vihvelin uses 'determinism' in a standard way, using it to represent "the thesis that a complete description of the state of the world at any time *t* and a complete statement of the laws of nature together entail every truth about the world at every time later than *t*" (Vihvelin 2011).

compatibilism is “unproblematic”, emphasizing that her definition allows the compatibilist to be agnostic about the truth of the determinism at the actual world (2011). However, given her conception of compatibilism, Vihvelin recognizes that it would lead to counter-intuitive results if she were to define ‘incompatibilism’ as the mere denial of compatibilism:

> Suppose, as some philosophers have argued, that we lack free will because free will is conceptually or metaphysically impossible, at least for non-godlike creatures like us (C.D. Broad 1934, G. Strawson 1986, 1994, 2002). If these philosophers are right, there are no free will worlds [i.e., possible worlds at which VFT is true]. And if there are no free will worlds, it follows that there are no deterministic free will worlds [i.e., possible worlds at which VFT and determinism are both true]. So if free will is conceptually or metaphysically impossible, at least for creatures like us, it follows that incompatibilism (as we have just defined it) is true. But this doesn't seem right. If it is conceptually or metaphysically impossible for us to have free will, then we lack free will *regardless* of whether determinism is true or false. And if that is so, then the incompatibilist cannot say the kind of things she has traditionally wanted to say: that the truth or falsity of determinism is *relevant* to the question of whether or not we have free will, that if determinism were true, then we would lack free will *because* determinism is true, and so on. (2011; emphasis in original)

In other words, given the way that Vihvelin defines ‘compatibilism’, if ‘incompatibilism’ were defined as the mere denial of compatibilism, everyone who denies Possibilism would qualify as an incompatibilist—even those who believe that *determinism poses no threat to free will*.

So that the mere denial of Possibilism does not qualify as an expression of incompatibilism, Vihvelin concludes that we must accept that incompatibilism is an expression of Possibilism. As such, if Impossibilism is true, then all expressions of Possibilism—which, according to Vihvelin, includes *both* compatibilism and incompatibilism—are false.

3. Two Counterexamples

On Vihvelin’s account, compatibilism and incompatibilism are mere contraries: both views may be false (because both are expressions of Possibilism), but only one may be true (because they forward logically inconsistent claims about the compatibility of free will and determinism). Clearly, casting compatibilism and incompatibilism as mere contraries (rather than contradictories) runs counter to the traditional view of the logical relationship between these two views. Now, as it happens, I believe that Vihvelin is right about this aspect of the relationship between compatibilism and incompatibilism and my reasons for agreeing with Vihvelin on this point will become apparent below. However, I reject Vihvelin's claim that we must distinguish incompatibilism from Impossibilism by casting incompatibilism as an expression of Possibilism.

Michael McKenna has criticized Vihvelin's characterization of incompatibilism, (correctly) saying that Vihvelin's "requirements for incompatibilism are too demanding" (2010: 432-33). As McKenna points out, there is logical space for a philosopher to endorse both incompatibilism and Impossibilism (2010: 433). McKenna describes a philosopher, let us call him "Moe", who believes that determinism precludes freedom and (for different reasons) that indeterminism precludes freedom as well.[5] Intuitively, Moe is an incompatibilist, but Vihvelin's Three-fold Classification entails that Moe is not an incompatibilist with respect to free will and determinism simply because he is *also* an incompatibilist about free will and *indeterminism*.

Given that the Three-fold Classification was not retired after McKenna's critique, let us consider an even more obvious example of an incompatibilist-impossibilist. Let us imagine a philosopher, call him "Max", who holds that determinism is necessarily true—perhaps because he is a law necessitarian who thinks that the actual laws are deterministic and/or because he

[5] There are practicing philosophers who hold Moe's views (cf. Colin McGinn (1993: 80) and Robert Nozick (1981: 37)). One might mistakenly think that Moe's views are an expression of what Derk Pereboom calls "hard incompatibilism", but hard incompatibilism is not an expression of impossibilism (cf. Pereboom 2009: 22). Notably, though, Vihvelin's definition of 'incompatibilism' has the odd implication that Pereboom would cease to qualify as an incompatibilist if he were to give up his view that agent causation is coherent and, so, endorse impossibilism—McKenna also points to this oddity in his critique of Vihvelin's definition (McKenna 2010: 433, fn. 7).

believes that the notion of “indeterministic laws” is incoherent.[6] Max also believes that compatibilism is false because he believes that necessarily, determinism precludes free will. In short, Max believes: (1) there is no metaphysically possible world at which the conjunction of determinism and VFT is true *because* necessarily, determinism precludes free will, (2) there is no metaphysically possible world at which determinism is false, and, so, (3) there is no metaphysically possible world at which someone acts freely. Applying Vihvelin’s criteria, Max is not an incompatibilist because he holds that determinism precludes free will in *every* possible world rather than a mere *subset* of all metaphysically possible worlds. Vihvelin’s characterization of incompatibilism now seems indefensible.

Even in the face of such compelling counterexamples, though, one might believe that the unintuitive implications of the Three-fold Classification must be accepted because there could be no *superior* mapping of the logical space occupied by (in)compatibilism and (im)possibilism. However, I will demonstrate that there is at least one superior classification schema available and show, thereby, that these counterexamples speak decisively against the Three-fold Classification.

4. Incompatibilism

Vihvelin begins her project with an assumption about the correct view of compatibilism. I find that the proper characterization of incompatibilism is much less contentious. Incompatibilism is, roughly, the view that it is metaphysically impossible for a human-like being to be both free and determined *because* necessarily, determinism undermines free will. More formally, we can express this central tenet of incompatibilism as follows:

[6] There is common agreement that the laws of nature of either deterministic or indeterministic, but there might be logical space for both determinism and indeterminism to be false at a possible world if there are no natural laws at that world. To avoid problems related to this issue, we might add that Moe and Max think that the notion of a non-law-governed world is also incoherent, or at least that each thinks that a viable candidate for free action could not exist at such a world. Indeed, in my view, it is hard to imagine what, if not some set of laws, could account for the perdurance of an object like a rock, let alone the perdurance of a cognitively sophisticated being with the type of knowledge and history of reflective self-awareness that is part and parcel of being a candidate for free agency. (My thanks to Robert Rupert for this suggestion.)

Using the terms introduced above and where
'Dxy' represents *x is determined by the physical laws of nature to perform y*
'bc' represents explanatory "because"[7]
'□' represents strong metaphysical necessity

The Strict Incompatibility Thesis (I): It is strongly metaphysically necessary that anyone who is determined to perform an action is someone who does not freely perform that action *because her action is determined*;
$(I) =_{df} \Box\forall x\forall y((Hx \,\&\, Ax \,\&\, Dxy) \rightarrow ({\sim}Fxy\text{-}bc\text{-}Dxy))$.[8]

Incompatibilism may have other defining tenets, but (I)—or some very similar thesis—is certainly among them.

Earlier, I was critical of Vihvelin's claim that compatibilists and incompatibilists must endorse ◊VFT. However, I agree that there is an (in)compatibility-neutral thesis which must be endorsed by the incompatibilists. First, someone who holds that determinism is a threat to free will cannot plausibly hold that determinism is necessarily false. To say that determinism is necessarily false is to say that it is impossible that determinism undermines free will. That is, if there is no possible world at which determinism is true, then the proposition that someone performs an act that is not free because the act was determined is false at every possible world. Likewise, if someone were to deny that there is a possible world at which human-like beings

[7] I am using the explanatory as opposed to a causal, evidential or inferential sort of 'because' here. The fact that 'because' is not a truth-functional connective does not mean that there is something ill-formed about this statement of incompatibilism, nor does it imply that there is no fact of the matter whether this statement of incompatibilism is true or false. The simple sentence "A because B" is true if and only if A and B are each true and (all things being equal) the truth of B provides a sufficient explanation for the truth of A.

[8] Some readers may wonder how there could be a formal proof of the entailment claims I make in this paper given the connective role of the explanatory 'because' in (I) and (C) (for a statement of the latter, see Section 6). The answer is that even though 'because' is not a truth-functional operator, there is a partial truth-table for 'because' which can be used to support these entailment claims. Just as a true conjunction (*A* and *B*) entails that each conjunct is true, what we might call the "bejunction" (*A* because *B*) entails that each "bejunct", bejunct *A* and bejunct *B*, is true. Unlike a conjunction, however, the inverse is not true: the truth of each bejunct is not sufficient for the truth of the bejunction. In addition to the truth of each bejunct, the truth of a bejunction requires that the propositions expressed by each bejunct stand in a special relationship—the explanatory 'because' relationship (described in fn. 7). Thus, a truth-value must be assigned to a bejunction as a whole based on whether its bejuncts are true *and* whether these bejuncts are related in the correct way. In other words, a bejunction must be treated like a simple sentence. However, based on the stable logical properties of 'because', we can use the following formula to extract information from the bejunction in a formal proof: ***TRANS***: $\Box\forall x\forall y\,(({\sim}Fxy\text{-}bc\text{-}Dxy)\rightarrow{\sim}Fxy)$. Using TRANS and the rules of S5, one can generate a formal proof which demonstrates the truth of each of the entailment claims made in this essay.

exist, or even if he were to deny that human-like beings exist at one of those possible worlds at which determinism is true, this person would undercut the incompatibilists' proposal that the truth of determinism explains why no human-like beings are free at worlds where determinism is true—beings that do not exist do not perform actions and, so, *a fortiori* do not perform *free* action. In short, incompatibilism seems to presuppose that the following thesis is at least metaphysically possibly true:

> *The Determined Human-like Being Thesis* (DBT): The conjunction of the thesis of determinism and the proposition that some human-like being performs an action; (DBT)$=_{df}$ $\exists x \exists y(Hx \,\&\, Ay \,\&\, Dxy)$.[9]

If my claim that incompatibilists must endorse ◊DBT is not obviously true, it is still significantly less controversial than Vihvelin's similar claim about ◊VFT. In the light of the stories of Moe and Max, Vihvelin's claim that incompatibilists *must* endorse ◊VFT is apparently false. At the very least, I believe that most (if not all) self-identifying incompatibilists would agree that ◊DBT is true—after all, presumably *we* are human-like beings and the view that determinism is incoherent is rarely mentioned, let alone defended. So, hereafter, I will use "Incompatibilism" to refer to the view that the conjunction of (I) and ◊DBT is true.

Notably, ◊DBT is implied by the (widely accepted) definition of 'compatibilism' that Vihvelin endorses. So, if we were to adopt Vihvelin's definition of 'compatibilism' and my preferred definition of 'incompatibilism', it would follow that both compatibilists and incompatibilists must endorse ◊DBT. In other words, the negation of ◊DBT would imply that *both* compatibilism and incompatibilism are false. Now, if accepting that compatibilism and incompatibilism are mere contraries seems to be an unacceptably high price for the benefits of my schema, recall that the Three-fold Classification which my schema is designed to supplant

[9] If determinism (as defined in fn. 4) is true, then any action performed by a human-like being is an action that he or she is determined by the laws to perform: $\square(\text{Determinism} \rightarrow \forall x \forall y(Hx \,\&\, Ay) \rightarrow (Dxy))$.

also presents compatibilism and incompatibilism as mere contraries. As such, the alternative mapping I have suggested is not *inferior* to Vihvelin's schema in this regard, even though I imagine that many (think that they) would *prefer* a mapping on which compatibilism and incompatibilism are contradictories rather than contraries.

More importantly, though, I will argue below (in Section 6) that we must reject Vihvelin's preferred definition of 'compatibilism'. I will suggest two superior characterizations of compatibilism and argue that there is at least one way of refining the definition of 'compatibilism' that would reflect the standard view that a philosopher cannot deny the truth of both compatibilism and incompatibilism—which is notable, for (to my knowledge) *no* characterization of compatibilism and incompatibilism in the free-will literature does this.

5. (Im)possibilism and (In)compossibilism

In order to make the logical relationships between Incompatibilism, Possibilism, and Impossibilism more transparent, let us revisit the formal definitions of the latter two views. As suggested above, Possibilism and Impossibilism can be understood in terms of the following theses:

The Free-Will Possibility Thesis ($\Diamond$VFT) $=_{df}$ $\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Fxy)$

The Free-Will Impossibility Thesis ($\sim\Diamond$VFT) $=_{df}$ $\sim\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Fxy)$[10]

Possibilism is the view that the Free-will Possibility Thesis is true, while Impossibilism is the view that the Free-will Impossibility Thesis is true.

[10] As discussed above (in fn. 3), these are *qualified* expressions of possibilism and impossibilism. The reader should read "Possibilism" as shorthand for "Possibilism Regarding Human-like Beings" and the same goes, *mutatis mutandis*, for "Impossibilism". Where 'Sx' represents *x is an entity*, "Unrestricted Possibilism" can be expressed by the formula '$\Diamond\exists x\exists y(Sx \,\&\, Ay \,\&\, Fxy)$' and "Unrestricted Impossibilism" can be expressed by '$\sim\Diamond\exists x\exists y(Sx \,\&\, Ay \,\&\, Fxy)$'. (Notably, in her initial characterization of impossibilism, Vihvelin toyed with the idea of using the term 'impossibilism' for the latter view (Vihvelin 2008: 303).)

As Vihvelin's characterization of compatibilism reflects, compatibilism is commonly taken to be an expression of Possibilism. According to Vihvelin, compatibilism is logically equivalent to a view that I will call "Compossibilism", where the latter is the view that the following thesis is true:

> *The Compossibility Thesis* (P): It is strongly metaphysically possible that determinism is true and there exists a human-like being who is free in performing some action even though he is determined to perform that action;
> (P) $=_{df} \Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$.

Let use the name "Incompossibilism" for the view that the Compossibility Thesis is false. Upon review, we observe that Incompatibilism is nothing more than Incompossibilism defended in a particular way. Alternatively, we could say that Incompossibilism is Incompatibilism's negative thesis, while the diagnostic element of (I) expresses Incompatibilism's positive thesis. More generally: Incompatibilism, Incompossibilism, and Impossibilism each independently entail the negation of Compossibilism, and the latter entails the negation of each of the former three views.[11]

Notably, my schema also leaves space for another interesting view. Imagine a philosopher, call him "Bud", who endorses Impossibilism based solely on the Basic Argument (an argument which purportedly shows that it is metaphysically impossible for *any* type of being (god-like or not) to have free will because genuine freedom would require that one be a *causa sui* (c.f. Strawson 1994)).[12] Now, as a proponent of the Basic Argument, Bud would be committed to the view that there is no metaphysically possible being that both wears a green shirt and

[11] Although (P) is not the mere denial of (I), (P) *entails* the denial of (I) and, so, the denial Incompatibilism. Alternatively, if Incompatibilism is true, then (P) must be false—for it follows from the conjunction of (I) and the first three conjuncts of (P) that ~(P).

[12] In other words (drawing from fn. 10), the Basic Argument supports *both* "Impossibilism Regarding Human-like Beings" and unqualified "Impossibilism". The latter version of impossibilism entails the former, but not vice versa. The reader should be aware, then, that the type of impossibilism which is supported by the Basic Argument is a more sweeping type of impossibilism than the view Vihvelin describes under the name "Impossibilism".

performs a free action. Of course, this does imply that Bud believes that there possibly exists some agent who is not free *because* the agent is wearing a green shirt. Likewise, Bud would agree that there is no agent who lives in a deterministic universe and performs a free action, but he would *reject* the Incompatibilist's claim that the obtaining of deterministic laws *explains* why there are no free actions in such a universe. Put another way, according to Bud, the reason that there is no possible world at which the conjunction of determinism and VFT is true is simply that VFT is necessarily false. Since VFT is equivalent to the final conjunct in (P), the conclusion of the Basic Argument entails the negation of (P), i.e., the Basic Argument entails Incompossibilism. Hence, Bud is an Impossibilist who endorses Incompossibilism but rejects Incompatibilism—Bud is a "Non-Incompatibilist-Impossibilist".[13]

We now have a thorough mapping of the logical space occupied by (im)possibilism, (in)compossibilism, and incompatibilism. Summarizing the major logical relationships:

- *Possibilism* and *Impossibilism* are contradictory views;
- *Compossibilism* and *Incompossibilism* are contradictory views;
- *Compossibilism* and *Incompatibilism* are merely contrary views;
- *Incompatibilism* entails but is not entailed by *Incompossibilism*;
- *Incompossibilism* is entailed by but does not entail *Impossibilism*;
- *Incompatibilism* and *Impossibilism* are logically consistent and logically independent views.

Given these logical relationships, one can do any of the following consistently: (a) endorse both Possibilism and Incompatibilism (as free-will libertarians do), (b) remain agnostic about Possibilism yet endorse Incompatibilism,[14] (c) reject Possibilism and endorse Incompatibilism (like Moe and Max), or (d) reject both Possibilism and Incompatibilism (like Carl). All that

[13] Joseph Keim Campbell holds that all impossibilists are incompatibilists, saying that "if free will is metaphysically impossible, it cannot co-exist with anything; ipso facto, it cannot co-exist with determinism and incompatibilism is true" (2011: 54). I contend that Campbell (mis)conceives of incompatibilism as a mere rejection of Compossibilism, which is to say that Campbell fails to recognize the distinction between mere Incompossibilism and Incompatibilism.

[14] Derk Pereboom seems to hold this view (see fn. 5).

remains of my positive project is to sketch out how each of these major views relates to compatibilism.

6. Compatibilism

While Vihvelin considers her definition of 'compatibilism' to be uncontroversial, I deny that there is any orthodox or uncontroversial way of expressing compatibilism. Still, *pace* Vihvelin and others who interpret "compatibility" as mere "compossibility", we can be certain that compatibilism is *not* logically equivalent to Compossibilism. Indeed, the truth of this non-identity claim is quickly demonstrated: (P) is logically consistent with the view that there exists, at some metaphysically possible world, a human-like being who performs an action that is not free *just because this action is determined by the natural laws.* In other words, (P) does not entail the negation of ◊∃x∃y (Hx & Ay & Dxy & (~Fxy-bc-Dxy)—but (intuitively) *compatibilism* does entail the negation of the latter existential claim. Since Compossibilism is expressed fully by the claim that (P) is true but compatibilism is not, it follows that Compossibilism and compatibilism are not logically equivalent.

Additional support for my claim that compatibilism, as commonly understood, is not equivalent to Compossibilism can be drawn from a review of the familiar compatibilist position known as "soft determinism". Typically, soft determinism is described as the view that determinism and VFT (or some similar free-will thesis) are both true (cf. van Inwagen 2008: 330; Vihvelin 2011; Kane 2002: 290). According to this portrayal, soft determinism is the view that the actual world is a possible world at which the conjunction of VFT and DBT is true—making soft determinism an instance of (P). However, let us review the description of soft determinism offered by William James in the passage where he gave this view its name:

> Nowadays, we have a *soft* determinism which abhors harsh words, and, repudiating fatality, necessity, and even predetermination, says that its real name

> is freedom; for freedom is only necessity understood, and bondage to the highest is identical with true freedom. (1956: 149)

In saying that the soft determinist is someone who holds that "freedom is only necessity understood" and suggesting that determinism is "identical with" and, so, inseparable from "true freedom", James seems to mean that the soft determinist holds that determinism presents *no threat whatsoever* to free will.

I believe that James intends for us to understand the soft determinist as someone who endorses (P), but also a principle like:

> *The Strict Compatibility Thesis* (C): It is strongly metaphysically impossible that there exists an agent who does not act freely *merely because her actions are determined*;
> $(C) =_{df} \Box\forall x\forall y((Hx \,\&\, Ay \,\&\, Dxy) \rightarrow {\sim}({\sim}Fxy\text{-bc-}Dxy))$.[15]

Simply put, (C) expresses the view that necessarily, determinism is not a threat to anyone's freedom and, so, the *incompatibilists are categorically wrong* when they assert that determinism is sufficient to undermine a person's freedom. I contend that (C)—or some very similar strict principle—is among the defining tenets of compatibilism. Let us refer to the view that the conjunction of (P) and (C) is true as "Compossibility-Compatibilism".

Compossibility-Compatibilism expresses the view that determinism is *in no way whatsoever* a threat to free will, which seems to be the general view that is most commonly associated with the term 'compatibilism'. Were we to accept that compatibilism is identical to Compossibility-Compatibilism, an appealing formal similarity between compatibilism and Incompatibilism would result. As noted above, Incompatibilism entails Incompossibilism. As such, Incompatibilism can be described as having both a *positive* incompatibilist thesis, the conjunction of (I) and $\Diamond$DBT, and a *negative* incompossibilist thesis, ~(P). The negative thesis of

[15] Alternatively: ${\sim}\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Dxy \,\&\, ({\sim}Fxy\text{-bc-}Dxy))$.

Incompatibilism is the contradictory of Compossibility-Compatibilism's positive thesis, i.e. (P); the positive thesis of Incompatibilism is the contrary of Compossibility-Compatibilism's negative thesis, i.e. (C). Given that Compossibility-Compatibilism seems to capture the view most commonly associated with 'compatibilism' and its theses are complements of the theses of (Incompossibility-)Incompatibilism, one may think it obvious that compatibilism is identical to Compossibility-Compatibilism.

Mulling over the place of compatibilism in the broader dialectic, though, we see that there is reason to doubt that compatibilists, *qua* being compatibilists, must endorse (P). If we were to accept that (P) is a defining tenet of compatibilism, we would also have to accept that any argument against (P) is an argument against compatibilism. The negation of (P) is entailed by both Incompossibilism and Impossibilism, and each of the latter two views might be true even if Incompatibilism is false. If we were to identify compatibilism with Compossibility-Compatibilism, then (P) would be a defining tenet of compatibilism. It follows, then, that *both* compatibilism and Incompatibilism could be false. As with Vihvelin's Three-fold Classification, then, the three-fold classification of (Compossibility-)Compatibilism, Incompossibilism, and Incompatibilism would leave a significant logical gap between *arguments against compatibilism* and *arguments for incompatibilism*.

For those who do not wish to accept the existence of a logical gap between compatibilism and incompatibilism, I hasten to point out one way to narrow this gap. One might argue that compatibilism would be best understood as the modest view that (C) is true—after all, one can hold that (C) is true without endorsing (P).[16] Seen this way, compatibilism is a strictly negative

[16] Or perhaps compatibilism should be understood as (C) taken together with certain (in)compatibility-neutral background assumptions. For instance, the conjunction of (C) and ◊DBT entails ~(I), while the conjunction of (I) and ◊DBT entails ~(C). This makes the conjunction of (C) and the (in)compatibility-neutral assumption ◊DBT

thesis that responds to the positive “because” claim made in (I). While (C) on its own does not entail the negation of (I), the conjunction of (C) and ◊DBT entails the negation of (I). Thus, the conjunction of (C) and Incompatibilism would entail a contradiction, so Incompatibilism and (C) cannot both be true. This means that if we were to accept that compatibilism is the view that (C) is true, even the Impossibilist who is a strict free-will error theorist, i.e. someone who holds that necessarily all first-order freedom claims are false, would have to accept either that Compatibilism or that Incompatibilism is true: the error-theorist Impossibilist could deny both (I) and (P), but even he could not hold that both (C) and (I) are false.[17]

Technically, though, even if we were to agree that (C) is the only defining tenet of compatibilism, compatibilism and Incompatibilism would not be contradictory views. ◊∃x∃y(Hx & Ay & Dxy & (~Fxy-bc-Dxy)) and ◊∃x∃y(Hx & Ay & Dxy & ~(Fxy-bc-Dxy)) are logically consistent views and if the conjunction of these views were true, then both (C) and (I) would be false. However, a question now arises: On what grounds could a philosopher *defend* the view that, all other things being equal, some actions are not free *solely in virtue of the fact they are determined by the natural laws* and yet other actions are free *despite being determined*? Upon reflection, there seems to be no *principled* way of defending the view that both ◊∃x∃y(Hx & Ay

appear to be a viable candidate for the correct expression of compatibilism. Whether compatibilism includes ◊DBT as a defining tenet is a highly contentious matter and I will not settle the matter here.

[17] Now, whether a strict free-will error theorist will endorse (C) or (I) will depend upon the particulars of his/her other views. For instance, recall Max, the Incompatibilist-Impossibilist discussed above. Let us now add that Max believes that necessarily, all first-order freedom claims are false. According to Max, (I), i.e., □∀x∀y((Hx & Ax & Dxy) → (~Fxy-bc-Dxy)), is true because Fxy is false at every possible world *because* Dxy is true at every possible world. Since Max endorses (I) and ◊DBT, he is an Incompatibilist despite being a strict free-will error theorist. By contrast, recall the story of Bud, the philosopher whose commitment to Impossibilism followed from his endorsement of the Basic Argument. As discussed above, Bud denies (I), and thereby denies Incompatibilism. If we now add that Bud is a strict free-will error theorist, his views also entail (C). That is, according to Bud, there is no possible world at which (1) the proposition *that the laws are deterministic* is true, (2) the proposition *that there exists some human-like being whom freely performs some action* is false, and (3) the truth of the former proposition provides a sufficient *explanation* for the falsity of the latter. In other words, Bud’s views entail that ~(~Fxy-bc-Dxy) is necessarily true—which is to say that Bud’s view imply that (C) is true because this thesis is a conditional which has a necessarily true consequent.

& Dxy & (~Fxy-bc-Dxy)) and ◊∃x∃y(Hx & Ay & Dxy & ~(Fxy-bc-Dxy)) are true.[18] The upshot here is that even though there is *technically* logical space for the view that both (I) and (C) are false, there does not seem be logical space to give a *rational defense* of this view.[19] So, by understanding compatibilism solely in terms of (C), we would accept that there is a narrow logical gap between compatibilism and incompatibilism and yet deny that there is a logical gap between *arguments for the conclusion that compatibilism is true* and *arguments for the conclusion that incompatibilism is false*.[20]

Of course, understanding compatibilism as the view that (C) is true would come with its fair share of uncomfortable implications. Above all, it would seem quite revisionary to define ‘compatibilism’ as the view that (C) is true when we consider that it would mean that we would have to accept the coherence of “*Compatibilist*-Impossibilism”, the view that the conjunction of (C) and ~◊VFT is true. I suspect that the debate over how we should use the term ‘compatibilism’ will boil down to the debate over whether it is preferable to accept (1) that there is logical space for both compatibilism and impossibilism to be true, or (2) that there is a logical

[18] Notably, Incompatibilism would be false if ◊DBT were false, but notice that the antecedent of (C) would be false if ◊DBT false. This means that (C) is true if ◊DBT is false. So, on the assumption that (C) exhaustively expresses compatibilism, compatibilism and Incompatibilism are both false only if ~(C) and ~(I).

[19] Although I cannot make the case here, I think that a similar strategy could be used to reply to Seth Shabo’s claim that there can be “good and interesting incompatibilist arguments” which are not arguments for incompatibilism (2011: 370). As characterized in this essay, the traditional free-will (in)compatibility debate is over whether necessarily, determinism is *sufficient* to undermine free will. Shabo, by contrast, seems to be picking up on a parallel quasi-(in)compatibility debate (going back at least to Warfield 2000) over whether the truth of determinism at a world *W*, when taken together with certain other contingently true propositions at *W*, jointly entail that some or all of the people in *W* lack free will. Given that “incompatibilist” arguments of the latter kind fall short of arguments for (I) and threaten *neither* (C) nor (P), it seems that Shabo is misusing the adjective “incompatibilist”. At the very least, we might introduce a technical distinction between “incompatibilist” and “incompatibilistic” arguments (much like the distinction between “Hellenic” and “Hellenistic” philosophy), using the former to apply to arguments which support Incompatibilism and the latter to apply to those more modest arguments which support the quasi-incompatibilist view that possibly, determinism and some contingent proposition *P* jointly entail the proposition that a subject *S* lacks free will.

[20] In saying this, I do not deny that there is logical space for an argument in support of free-will non-cognitivist, roughly the view that all first-order freedom claims are *meaningless*. A free-will non-cognitivist might appeal to an argument like the Basic Argument in support of his view, but the free-will non-cognitivist cannot say that the argument shows (helps to show) that compatibilism and incompatibilism each *false*; he can say only that the Basic Argument reveals that neither compatibilism nor incompatibilism expresses a *meaningful* view.

gap between arguments against compatibilism and arguments for incompatibilism. However we ultimately define ‘compatibilism’, though, (C) is notable because this thesis allows us to see that there is logical space for a philosopher to take up an “anti-Incompatibilist” stance without committing to the truth of Compossibilism. So, whatever we choose to *call* the view that (C) is true, this is an interesting, independent “compatibilistic” view which has been overlooked until now.

If nothing else, the above discussion displays that the common understanding of compatibilism is quite muddled. As a result, the locus of the dispute between the “compatibilists” and incompatibilists is also far from clear.[21] Hence, *some* revisionary move must be made—either the definition of ‘compatibilism’ must be refined, the term should be used to refer to a *collection* of distinct compatibilistic theses, or the term must be jettisoned from the debate. Which of these options we should favor and, if ‘compatibilism’ is kept, which revised definition of this term we should endorse, are issues which strike me as worthy topics of debate. One might even say that a debate on this issue is long overdue—but it does not follow that I carry the

[21] For those who think that the taxonomical issues in this paper carry little philosophical import, consider the conclusion of Derk Pereboom’s famous “Four-Case Argument”. According to Pereboom, the Four-Case Argument reaches (in its second stage) the “incompatibilist” conclusion that there is no set of compatibilist-friendly sufficiency conditions for moral responsibility (2001: 112). In the light of the discussion in this paper, we can now see that Pereboom wrongly believes that his argument supports incompatibilism *in virtue of the fact that the argument undermines Compossibilism*. Likewise, in McKenna’s critique of Vihvelin, McKenna complains that Vihvelin failed to identify the Four-Case Argument as an argument for incompatibilism, even though McKenna there describes the Four-Case Argument as an argument against Compossibilism and not an argument for Incompatibilism (2010: 439). In other words, both Pereboom and McKenna fail to appreciate the distinction between arguments for mere Incompossibilism and arguments for Incompatibilism. (Admittedly, Pereboom provides a *best-explanation argument* which identifies determinism as a threat to free will, and, so, there is room to argue that the Four-Case Argument has all of the makings of an argument for Incompatibilism. Perhaps so, but neither Pereboom nor the many others working on the topic of manipulation arguments have recognized the logical gulf between Incompossibilism and Incompatibilism. As a result, the logical structure of the Four-Case Argument (and how the best-explanation argument fits into the argument overall) and the shared formal features of manipulation arguments more generally, are still unclear. (For a detailed discussion of the formal structure of the Four-Case Argument and other manipulation arguments, see my “Misimpressions of the Manipulation Argument”, [unpublished manuscript]).)

argumentative burden of settling these thorny debates here.[22] There is clearly *some way* of resolving these debates which is consistent with the taxonomy that I have started to develop, and this suffices to show that there are viable and superior alternatives to Vihvelin's Three-fold Classification.

7. Concluding Remarks

Since Vihvelin builds an idiosyncratic view of the dialectic between compatibilists and incompatibilists upon the firmament of the Three-fold Classification, my critique of the Three-fold Classification could easily be extended to Vihvelin's other views. In brief, Vihvelin claims that the incompatibilists carry a heavier argumentative burden than compatibilists because Vihvelin (wrongly) believes that incompatibilists (*qua* being incompatibilists) endorse Possibilism. Since Vihvelin conceives of compatibilism in terms of merely (P), she claims that the compatibilist and the incompatibilist each carry the argumentative burden of showing that there is some possible world at which her version of the free-will thesis (i.e., VFT) is true. Moreover, she contends that the incompatibilist carries the *extra* burden of showing that the free-will thesis (VFT) is true *only* at worlds where *indeterminism* is also true. As we have seen, though, the incompatibilist is not committed (*qua* being an incompatibilist) to Possibilism, which means that the incompatibilist carries *neither* of these argumentative burdens. Indeed, having seen that one can reject Incompatibilism without endorsing Possibilism (by holding that (C) is true) we have reason to wonder whether even the *compatibilist* carries the burden of defending Possibilism.

In sum, I have argued that Kadri Vihvelin's Three-Fold Classification mischaracterizes compatibilism, incompatibilism, the logical relationship between incompatibilism and

[22] I propose a solution to these debates and more in my "(In)compatibility" [unpublished manuscript]).

impossibilism, and, so, promotes a misguided view of the argumentative burdens carried by the proponents of each of these views. I have demonstrated that there is an alternative classification that allows us to (1) capture the modal force typically associated with incompatibilism, (2) block the entailment from impossibilism to incompatibilism, and yet (3) deny Vihvelin's (untenable) claim that incompatibilism entails possibilism. In short, I have shown that we need not accept Vihvelin's problematic Three-fold Classification in order to secure any of its touted benefits, so the Three-fold Classification schema *can* and *should* be rejected.

CHAPTER FIVE

(IN)COMPATIBILITY

1. Introduction

Compatibilism and incompatibilism are two of the most familiar views in the contemporary free-will debate, yet there is no adequate formal expression of either of these views available. As a result, the debates between proponents of compatibilism, incompatibilism, and other major views (like impossibilism) are widely mischaracterized and misunderstood. A comprehensive formal mapping of the logical space in which the free-will debate takes place is needed and is what this essay will provide.

According to Peter van Inwagen, ‘compatibilism’ is best defined as the thesis that the conjunction of the thesis of determinism and the free-will thesis (roughly the thesis that some agent like us exists who performs a free action) is true at some metaphysically possible world; ‘incompatibilism’ is best defined as the thesis that compatibilism is false (cf. van Inwagen 1983: 12, 2008: 330). Most philosophers who make an effort to give precise formal definitions of ‘compatiblism’ and ‘incompatibilism’ follow van Inwagen’s lead (cf. Mele 1995: 142; Campbell 2011: 21). However, as Kadri Vihvelin has pointed out, when we define ‘incompatibilism’ as a strictly negative thesis, i.e. as the mere denial of the thesis that the conjunction of the free-will thesis and the thesis of determinism is true at some metaphysically possible world, we are saddled with a counterintuitive view of the relationship between incompatibilism and impossibilism, i.e. the view that the free-will thesis is necessary false (cf. Vihvelin 2011, 2008). All impossibilists deny the possible truth of the conjunction of determinism and the free-will thesis because it follows from impossibilism that the latter conjunct is necessarily false. Thus, van Inwagen's preferred definition of 'incompatibilism' wrongly implies that even those impossibilists whom *deny* that determinism precludes free will are nonetheless incompatibilists.

In an effort to preserve the common definition of compatibilism and the intuitive distinction between impossibilism and incompatibilism, Vihvelin forwards what she calls a "Three-fold Classification" of compatibilism, incompatibilism, and impossibilism (2011; 2008). This Three-fold Classification is built around Vihvelin's attempt to block the entailment from impossibilism to incompatibilism, which she does by defining 'incompatibilism' and 'compatibilism' so that each is logically inconsistent with impossibilism. However, as Michael McKenna (2010: 432-33) and I (Chapter 4: "Beyond the Three-fold Classification") have argued, there is clearly logical space for the "incompatibilist-impossibilist", i.e. the philosopher who endorses incompatibilism and impossibilism. So, Vihvelin's proposed account of incompatibilism, too, fails to capture the intuitively correct logical relationship between incompatibilism and impossibilism.

My project begins with a brief overview of the key concepts and (in)compatibility-neutral presuppositions of the interlocutors in the debate over the "compatibility" of free will and determinism. Over the course of the next few sections (Sections 3-8), I discuss various modal theses that one might associate with compatibilism and incompatibilism. Emerging from this discussion is that the logical relationships between compatibilist and incompatibilist views are much more complicated than is commonly thought. Most notably, philosophers commonly conflate compatibilism and a view I call "compossibilism" and also conflate incompatibilism with a view that I call “incompossibilism”. While teasing out a clear definition of ‘incompatibilism’ is relatively straightforward, I cast doubt on the assumption that there is a single view that (uncontroversially) answers to the name “compatibilism”. Finally (in Section 7), I present and defend my preferred definition of ‘compatibilism’. I argue that compatibilism and incompatibilism must be understood as contrary views (rather than contradictories). While Vihvelin also holds that compatibilism and incompatibilism are mere contraries, her definition of ‘incompatibilism’ is false and her definition of ‘compatibilism’ is incomplete. In addition to these profound errors, Vihvelin’s schema also allows that there can be arguments against compatibilism which are not arguments for incompatibilism, but my preferred characterization leaves no such logical gap.

I close (in Section 7) with a discussion of how working with impoverished views of compatibilism and incompatibilism can inhibit progress in the free-will debate. As a case in point, I discuss Michael McKenna’s “Manipulation Argument” and its most famous instance, Derk Pereboom’s famous “Four-Case Argument”. I argue that the Four-Case Argument, even if sound, does not rise to its billing as an argument for incompatibilism and McKenna’s template, which is inspired by the Four-Case Argument, does not outline an argument for incompatibilism. I do not deny that there are some manipulation arguments which provide a defense of incompatibilism, but I do deny that the literature on manipulation arguments provides and clear sense of what *makes* a given manipulation argument an argument for incompatibilism.Finally, I acknowledge that my preferred definition of compatibilism seems revisionary, but argue that is much less revisionary than it appears *prima facie*. In short, I argue that my preferred definitions are the first to reflect the strict (in)compatibility principles that have always been associated with compatibilism and incompatibilism but have never before been articulated adequately.

2. The Primary Free-Will (In)compatibility Debate

In contemporary literature, there are several distinct free-will debates. What I have to say in this essay will apply directly to at least two of them, what I call the "Primary Free-Will (In)compatibility Debate" and the "Secondary Free-Will (In)compatibility Debate".[1] The Primary Free-Will Debate centers on the (in)compatibility of *free will* and *determinism*, while the Secondary Free-Will Debate centers on the (in)compatibility of *free will* and *indeterminism*.

The formal structures of the Primary and Secondary (In)compatibility Debates are the same, so what I have to say below about the formal elements of the former can be extended, *mutatis mutandis*, to reveal the formal elements of the latter. Although there is a compatibilist solution and an incompatibilist solution to each of these (in)compatibility debates, I will use the terms 'compatibilism' and 'incompatibilism' as they are traditionally used, i.e. as candidate solutions to the Primary Free-Will (In)compatibility Debate. More specifically, I will use 'compatibilism' as shorthand for "Primary Free-Will Compatibilism" and 'incompatibilism' "Primary Free-Will Incompatibilism". While I will not discuss the Secondary (In)compatibility Debate in detail, allusion to this debate will be unavoidable when discussing some of the most famous instances of compatibilism and incompatibilism.

There are two concepts at the heart of the Primary (In)compatibility Debate: *free will* and *determinism*. Along standard lines, let us use 'determinism' as shorthand for:

> *The Thesis of Determinism* (TD) $=_{df}$ the thesis that the conjunction of a proposition P which expresses a past state of a universe *u* and a proposition L which expresses the natural laws of *u* entails any proposition P* which expresses a future state of *u*.

If TD is true, then the laws of our universe are deterministic. There are more precise ways of stating the thesis of determinism (cf. van Inwagen 1983: 65, 83), but TD suffices as a generic

[1] There are also Primary and Secondary Moral Responsibility (In)compatibility Debates (the debates over whether *moral responsibility* is (in)compatible with determinism and indeterminism, respectively) which presumably have the same formal structure, but I will discuss neither of these debates in this essay.

example of how the thesis of natural law determinism should be expressed. Indeed, a distinct debate will be required to reach agreement on the correct formulation of TD, i.e. the formulation of TD which should be used for purposes of the Primary Free-Will (In)compatibility Debate.[2] No thesis in this essay depends on the specific content of TD; the reader who disapproves of my statement of determinism should think of 'TD' as a placeholder for the ideal expression of natural law determinism.

The central point of contention in the Primary Free-Will (In)compatibility Debate is whether the truth of TD is consistent with the truth of what is typically called the "free-will thesis". Unfortunately, there is no standard statement of the free-will thesis. In one statement of this thesis, van Inwagen expresses the thesis as follows:

> The free-will thesis is the thesis that *we* are sometimes in the following position with respect to a contemplated future act: we simultaneously have both the following abilities: the ability to perform that act and the ability to refrain from performing that act […]. (2008: 329; my emphasis)

While I agree that most interlocutors in the free-will debate are concerned about whether or not *we* are free, I do not think that an adequate statement of the free-will thesis will appeal specifically to us. In my view, there is only one correct account of free will—maybe humans can have it, maybe only God can have it, maybe it is the other way around, or maybe no one can have it—and it is irrelevant to the correct definition of 'free will' whether *we* humans (sometimes) have it. Moreover, some philosophers (like me) are interested in the question of whether it is metaphysically possible for non-human beings to act freely, but van Inwagen's proposed free-will thesis is not sufficiently general to be useful in the broader debate about possibility of what is sometimes called "metaphysical freedom".

[2] For instance, someone who believes that miracles may occur without making a law false will not endorse the standard statement of TD I have provided.

Furthermore, van Inwagen's version of the free-will thesis seems biased against the view that free will is best understood in terms of sourcehood and *not* in terms of access to alternate possibilities. Thus, I recommend the following free-will thesis:

> *The Definitional Free-Will Thesis* (D-FWT): the thesis that an individual is sometimes in the following position with respect to a contemplated future act *y*: whether or not that being performs *y* is *up to* that individual.

Now, I do not claim that D-FWT presents the correct definition of "free will". In the first place, my statement of D-FWT is far too vague to be a proper definition. What we have in D-FWT is, however, a suitable *working* definition, i.e. one which is sufficiently non-committal that an interlocutor in the debate could accept it. As with determinism, an independent debate will be required to settle the matter of what it means for an action to be "up to" an individual in the right sort of way. Also, as with the thesis of determinism, my purpose in articulating a complete statement of D-FWT is to illuminate the general formal characteristics of a proper expression of this thesis. The key points in this essay in no way depend on how we understand free will; my statement 'E-FWT' can be viewed as a placeholder for the ideal statement of the conditions under which an individual performs an action which is "up to" the individual in right sort of way for acting of one's own free will.

The correct content of D-FWT is a highly contentious matter and there is a pragmatic wisdom in doing what one can to avoid unnecessary digressions into controversy whenever one easily can do so. In those contexts where having the correct definition of "free will" is not essential—such as in a discussion of the merely formal elements of the Primary Free-Will (In)compatibility Debate, as taken up here—I believe that we can avoid problematic allusions to a controversial or an overly generic definition of 'free will' by using a version of the free-will thesis which simply posits the existence of free agents. Indeed, in one of van Inwagen's early statements of the free-will thesis, he describes the free-will thesis as "the thesis that we *have* free will" (1983: 13-14; my emphasis). Clearly, this statement of the free-will thesis is not intended as a definition of 'free will' (for it would offer little more than the definiendum as the definiens); this thesis is about the *existence* of free agents and it presupposes that we know what 'free will' means.

As with van Inwagen's preferred definitional free-will thesis, though, his existential version of the free-will thesis is problematic insofar as it includes explicit reference to *us*. Following van Inwagen, Vihvelin's preferred statement of the free-will thesis is "the thesis that at least one *non-godlike creature* has free will" (Vihvelin 2011; my emphasis). Vihvelin's qualification is less severe than van Inwagen's, and, so, is a step in the right direction. However, there are philosophers participating in the Primary Free-Will (In)compatibility Debate whose views entail the denial of a more sweeping existential free-will thesis. For example, Galen Strawson is famous for offering a concise formulation of an oft-repeated argument in the history of the free will debate, what he calls "The Basic Argument" (cf. Strawson 1986). According to this argument, it is strongly metaphysically impossible for *anyone* or *anything* to have free will and/or moral responsibility because free will requires that one be a *causa sui*—and nothing can be a *causa sui*, not even God.[3] Clearly, then, philosophers who believe that the Basic Argument is sound will deny the truth of an even stronger existential free-will thesis than we get from either van Inwagen or Vihvelin.

I recommend that we express the existential free-will thesis as a maximally generic thesis, along the lines of the existential free-will thesis offered by Joseph Keim Campbell: "Someone has free will" (2011: 1). We can then place restrictions on this general thesis, as

[3] Put another way:"[T]rue self-determination is logically impossible because it requires the actual completion of an infinite regress of choices of principles of choice" (Strawson 1986: 29)

needed, to express the existential free-will theses that philosophers might wish to endorse. Formally, let us express Campbell's general version of the existential free-will thesis as follows:

> 'Sx' represents *x is an entity*
> 'Ay' represents *y is an action*
> 'Fxy' represents *x freely performs y (as characterized by FWT)*
>
> *The Existential Free-will Thesis* (E-FWT): There exists some entity *x* who freely performs some action *y*, where the relevant notion of 'free' is explicated in FWT; (E-FWT) $=_{df}$ $\exists x \exists y(Sx \ \& \ Ay \ \& \ Fxy)$.[4]

In this essay, I will use "Possibilism" to name the thesis that E-FWT is true at some metaphysically possible world and use "Impossibilism" to refer to the thesis that possibilism is false.[5] Alternatively, Possibilism and Impossibilism can be understood in terms of the following theses:

> *The Free-Will Possibility Thesis* ($\Diamond$E-FWT) $=_{df}$ $\Diamond\exists x \exists y(Sx \ \& \ Ay \ \& \ Fxy)$
>
> *The Free-Will Impossibility Thesis* ($\sim\Diamond$E-FWT) $=_{df}$ $\sim\Diamond\exists x \exists y(Sx \ \& \ Ay \ \& \ Fxy)$

Possibilism is the view that the Free-will Possibility Thesis is true, while Impossibilism is the view that the Free-will Impossibility Thesis is true.[6] *Pace* Vihvelin, I contend that neither compatibilists nor incompatibilists must be Possibilists, and I shall defend this view in detail below.

[4] In those contexts where one must forward a version one's preferred version of FWT (as one must if one is arguing that we do or do not have free will for some reason or other), 'possibilism' and 'impossibilism' may be defined in terms of FWT, equating the former with $\Diamond$FWT and the latter with $\sim\Diamond$FWT.

[5] The definition of 'Impossibilism' I give here breaks slightly from the most recent definition given by Vihvelin, who coined the term. Vihvelin (2011) describes the impossibilist as someone who believes merely that it is metaphysically impossible *for beings like us* (i.e. non-godlike beings) to have free will. However, this seems to present impossibilism as a weaker thesis that it really is. For instance, Vihvelin appeals to G. Strawson's Basic Argument (discussed above) as an argument for impossibilism. The Basic Argument concludes that, without qualification, free action is metaphysically impossible—a more sweeping impossibilism than Vihvelin describes. Thus, I will follow Vihvelin's original description of the impossibilist as someone who believes that "free will is metaphysically impossible" (2008: 304).

[6] Equivalently, ($\sim\Diamond$E-FWT) $=_{df}$ $\Box\sim\exists x \exists y(Sx \ \& \ Ay \ \& \ Fxy)$.

Although maximally generic statements of D-FWT and E-FWT have their place, I think that Vihvelin and van Inwagen are right insofar as they recognized that the Primary Free-Will (In)compatibility Debate is a debate about beings like us, i.e. cognitively sophisticated beings who, for all that, cannot perform miracles with respect to the natural laws which govern our universe. Thus, we also need an extistential free-will thesis which will allow us to focus our discussion on the freedom of beings like us without *defining* free will in terms of beings like us. For purposes of the Primary Free-Will Debate, then, let us use the following qualified version of the E-FWT:

> *The Existential Free-will for Human-like Beings Thesis* (E-FWT$_H$): There exists some human-liked being *x* who freely performs some action *y* (where the relevant notion of 'free' is explicated in FWT);
>
> (E-FWT$_H$): $\exists x \exists y(Hx \,\&\, Ay \,\&\, Fxy)$.

Notably, a philosopher may endorse E-FWT *without* endorsing E-FWT$_H$ (while E-FWT$_H$ entails E-FWT, the converse is not true), but D-FWT articulates the relevant notion of freedom in both existential free-will theses. Insofar as compatibilists and incompatibilists disagree about what free will is (e.g., whether freedom requires the actual-sequence ability to do otherwise or just a counterfactual ability to do otherwise), they disagree about the correct statement of D-FWT. Insofar as compatibilists and incompatibilists disagree about the "compatibility" of free will and determinism, they disagree about the possible truth of the conjunction of TD and E-FWT$_H$.

Notably, qualifying E-FWT to generate E-FWT$_H$ also leads to qualified versions of possibilism and impossibilism, which can be defined in terms of the following theses:

> *The Qualified Free-Will Possibility Thesis:*
>
> ($\Diamond$E-FWT$_H$) $=_{df}$ $\Diamond\exists x \exists y(Hx \,\&\, Ay \,\&\, Fxy)$

The Qualified Free-Will Impossibility Thesis:

$(\sim\Diamond \text{E-FWT}_{\text{H}}) =_{\text{df}} \sim\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Fxy)$[7]

"Possibilism-$_{\text{H}}$" is the qualified possibilist view that the Qualified Existential Free-will Thesis is true, and "Impossibilism-$_{\text{H}}$" entails and is entailed by the negation of Possibilism-$_{\text{H}}$.

I shall argue below that compatibilists and incompatibilists may reject both $\Diamond\text{E-FWT}_{\text{H}}$ and ◊E-FWT. However, there is one view that we *must* reject in order to preserve the coherence of the Free-Will (In)compatibility Debates: free-will non-cognitivism, i.e. the view that free-will sentences do not describe propositions and therefore can be *neither* true nor false. If free-will non-cognitivism is true, then the claims made by the interlocutors in both the Primary and Secondary Free-Will Debate are meaningless and the debate literally amounts to 2000-plus years of sound and fury, signifying nothing. Thus, if non-cognitivism is true, the distinctions that I attempt to draw in this essay will all be distinctions without a difference and none of the views here discussed are genuine candidate solutions to either the Primary or Secondary Free-Will (In)compatibility Debate. However, I contend that non-cognitivism is the *only* view that we must summarily rule out in order to proceed. The justification for this bold claim—bold because it implies that the coherence of the Free-Will (In)compatibility Debates do not require that either the concept *DETERMINISM* or the concept *FREE WILL* is coherent—will be become clear in the discussion below.

3. Compatibility and Compossibility

The technical term 'compatibilism' was first introduced sometime in the 1960s to refer to a view about the relationship between free will and determinism, but the term was quickly co-

[7] As discussed above (in fn. 5), these are *qualified* expression of possibilism and impossibilism. The reader should read "Possibilism" as shorthand for "Possibilism Regarding Human-like Beings" and the same goes, mutatis mutandis, for "Impossibilism". Where 'Sx' represents *x is an entity*, "Unrestricted Possibilism" can be expressed by the formula '◊∃x∃y(Sx & Ay & Fxy)' and "Unrestricted Impossibilism" can be expressed by '~◊∃x∃y(Sx & Ay & Fxy)'. (Notably, in her initial characterization of impossibilism, Vihvelin toyed with the idea of using the term 'impossibilism' for the latter view (Vihvelin 2008: 303).)

opted by those whose central interest is the compatibility of determinism and *moral responsibility*.[8] However, even when the term ‘compatibilism’ is used to express a thesis about the compatibility of free will and determinism—as opposed to adding (or substituting) an assertion about the compatibility of determinism and moral responsibility—the term is now used to express a surprising variety of views. On one hand, ‘compatibilism’ is a technical term, so different philosophers may define the term however they like. On the other hand, given the prominence of compatibilism as a position in the contemporary free-will debate, the number of fundamentally different views which go by the name “compatibilism” is striking and the lack of agreement about what unites these views as instances of compatibilism is troubling. In this section, I attempt to tease out the distinct theses that are most commonly associated with compatibilism.

A complete survey of the disparate definitions of ‘compatibilism’ would be quite an undertaking (and one that has been done for the most part, c.f. Doyle 2011). A small sampling of definitions of ‘compatibilism’ from some leading figures in the contemporary free-will literature will suffice to reveal the disparate uses of the term. As indicated above, van Inwagen defines ‘compatibilism’ as the view that (his preferred version of) the free-will thesis and determinism could both be true. Galen Strawson defines compatibilism in the same way as van Inwagen, but makes a point of stating that a compatibilist might hold “that [determinism] D is true, that D does not imply that we are unfree, but that it has not been shown whether or not we are free” (1986: 5). By contrast, Derk Pereboom, in his influential *Living Without Free Will*, uses the term ‘compatibilism’ to name the thesis that “whether or not determinism is true we have free will”,

[8] According to van Inwagen, Keith Lehrer coined the term ‘compatibilism’ (van Inwagen, "Moral Responsibility, Determinism, and the Ability to do Otherwise" 1999: 342, fn. 2). While Lehrer is unsure that it was he was officially the fist to use the term, he agrees with my claim that the term was not originally intended to apply to views about moral responsibility (in personal correspondence).

noting that the famous compatibilist Peter Strawson holds such a view (Pereboom 2001: xvi-xvii). Richard Double attributes both a positive and a negative claim to compatibilism, stating that compatibilism's positive claim is that "Under certain conditions, determined choices can be free" while compatibilism's negative claim is that "Undetermined choices cannot be free" (where the 'can' and 'cannot' express metaphysical (im)possibility) (1996: 101). Along similar lines, William James's 'soft determinism' (James 1956), according to which (among other things) determinism is true, we are free, and determinism is *required* for free will, is universally considered to be an expression of compatibilism.

I believe that van Inwagen's definition captures the view that philosophers most often associate with the term 'compatibilism', so this view seems to provide a natural starting point for our investigation of compatibility and compatibilism. Once again, according to van Inwagen, the compatibilist is someone who believes that determinism and the free-will thesis might both be true. In the language of possible worlds, van Inwagen's definition can be expressed as the thesis that there is some metaphysically possible world at which the laws are deterministic and some human-like being performs a free action. Or, using my terminology, van Inwagen's proposed definition of 'compatibilism' can be expressed as the thesis that the conjunction of TD and E-FWT_H is metaphysically possibly true. More concisely, we could say that on one common understanding of compatibilism, it is the view that the following thesis is true:

> *The (Primary) Compossibility Thesis* (P): At some metaphysically possible world, determinism is true and there exists a human-like being who freely performs some action even though he is determined to perform that action;
> (P) $=_{df} \Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$.

When we consider the set of all metaphysically possible human-like entities who live in a universe with deterministic laws, (P) is the claim that at least one of these entities performs a free action. Let us call the view that (P) is true “Primary Compossibilism”, or “Compossibilism” for short. Compossibilism is a familiar—and arguably the standard—expression of compatibilism.

Compossibilism is the positive thesis that is most commonly associated with compatibilism, but is compatibilism really identical to Compossiblism? In an attempt to answer this question, let us look more closely at James’s version of compatibilism: soft determinism. Typically, ‘soft determinism’ is defined as the view that TD and E-FWT$_H$ (or two very similar theses) are both true (cf. van Inwagen 2008: 330; Vihvelin 2011; Kane 2002: 290). This definition makes soft determinism seem like a mere commitment to the truth of the following thesis:

> *The Actual (Primary) Compossibility Thesis* (P$_A$): Determinism is true (at the actual world) and there exists some human-like being who performs a free action;
> (P$_A$) $=_{df}$ $\exists x \exists y(Hx \ \& \ Ay \ \& \ Dxy \ \& \ Fxy)$

Understood in this way, soft determinism is just an expression of Compossibilism.

Alternatively, though, van Inwagen has defined ‘soft determinism’ as the conjunction of determinism and compatibilism, as have many others (c.f., van Inwagen 1983: 13-14, Kane 2002: 290, and Doyle 2011: 424). Although I suspect that most compatibilists believe that E-FWT$_H$ is true (at the actual world), the compatibilist need not hold this view—and van Inwagen agrees (cf. 1983: 226, n. 14). This brings us back to G. Strawson’s comment (noted above) that the compatibilist might hold that determinism is true and that determinism does not imply that we are not free, yet remain agnostic about whether or not we are free. The compatibilist, might, for instance, believe that determinism is true, that determinism is compatible with free will, but worry that there exists some evil being which subjects all human-like beings to some type of freedom-undermining manipulation. If such a being exists, then no one at the actual world is free, but the fact that no free agents exist is unrelated to the fact that TD is true. As such, there may be some nearby world at which determinism is true and at which there exists no such malicious manipulator, and at that world there exists someone who performs a free action. While I doubt that any practicing compatibilist endorses precisely this odd view, the fact that there is logical space for such a view helps us to see that the mere conjunction of compatibilism and determinism does not entail (P_A).

Instead, it seems that the conjunction of determinism and compatibilism is best understood as a claim about a set of nomologically possible worlds. More specifically, given the way that van Inwagen and others understand compatibilism, the conjunction of determinism and compatibilism seems to entail (where '$\Diamond\!\!\!\!\!{\scriptscriptstyle N}\,$' represents nomological possibility):

> *The Nomological (Primary) Compossibility Thesis* (P_N): Determinism is true (at the actual world) and, at some possible world at which the laws of nature are the same as they are at the actual world (i.e. at some *nomologically possible* world) there exists some human-like being who performs a free action;
>
> $(P_N) =_{df} TD \ \& \ \Diamond\!\!\!\!\!{\scriptscriptstyle N}\,\exists x \ \exists y \ (Hx \ \& \ Ay \ \& \ Dxy \ \& \ Fxy)$

Soft determinism is the view that *both* TD and E-FWT_H are true (at the actual world), so soft determinism *entails* (P_N) but it is not *equivalent* to (P_N). However, since soft determinism is supposed to be a view about the actual world, it is clearly a mistake to define it as the conjunction of determinism and compatibilism.

I contend that soft determinism is much more than the view that (P_A) is true. In order to illuminate the defining tenets of soft determinism—and thereby shed light on the logical space occupied by compatibilism—let us look back to James's own presentation of the view. James, who coined the terms 'soft' and 'hard determinism' (c. 1884), describes soft determinism as follows:

> [D]eterminists today insist that they alone are freedom′s champions. Old-fashioned determinism was what we may call *hard* determinism. It did not shrink from such words as fatality, bondage of the will, necessitation, and the like. Nowadays, we have a *soft* determinism which abhors harsh words, and,

> repudiating fatality, necessity, and even predetermination, says that its real name is freedom; for freedom is only necessity understood, and bondage to the highest is identical with true freedom. (1956: 149)

Looking at James's own account of soft determinism, we can see that soft determinism asserts (P_A) (and so entails (P_N) and (P)), but also asserts something more than a mere compossibility claim. In saying that soft determinists believe that "they alone are freedom's champions", James describes the soft determinist as someone who believes that determinism is true and that we have free will, but also holds the view that free will is *incompatible* with *indeterminism*.[9] This thesis is equivalent to what Double labels "Compatibilism's Negative Claim" (mentioned above), or the thesis that "Undetermined choices cannot be free" (1996: 101). However, I contend that the thesis that undetermined choices cannot be free is in not, properly speaking, an expression of compatibilism.[10]

In order to see this, let us consider Double's Negative Claim as an independent thesis. Expressing the Negative Claim formally, we might state it as follows: $\Box\forall x\forall y((Hx \,\&\, Ay \,\&\, Fxy) \rightarrow Dxy)$. If that is right, then the Negative Claim does not entail any of the modal strains of Compossibilism detailed above; it does, however, entail $\sim\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Fxy \,\&\, \sim Dxy)$. The formula '$\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Fxy \,\&\, \sim Dxy)$' should be familiar, as it is identical to (P) but for the final conjunct in each formula: the final conjunct in (P) says the determinism is true, the final conjunct of the Negative Claim says that determinism is false. In my schema, the claim that $\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Fxy \,\&\, \sim Dxy)$ is a defining tenet of "*Secondary* Free-Will Compossibilism",

[9] Although commonly overlooked, others have recognized this feature of soft determinism (cf. G. Strawson (1986: 5) and Pereobom (2001:xvi)).

[10] Compatibilism is clearly a presupposition of the assertion "Free will requires determinism", as this assertion conversationally implies that free will *is* compatible with determinism, but compatibilism is not presupposed by the claim "Undetermined choices cannot be free" unless this claim is taken in conjunction with the claim "Free agency is metaphysically possible". As I discussed above, we should not consider the truth of the latter claim (understood as $\Diamond$E-FWT$_H$) to be a background assumption of the Primary Free Will Debate, as we must not artificially rule out the possibility that our investigation of free will shall reveal that no one could possibly be free ($\sim\Diamond$E-FWT).

a view in the *Secondary* Free-Will (In)compatibility Debate. Less formally, Secondary Free-Will Compossibilism is the view that there exists some free human-like being at some possible world at which determinism is false, i.e., a world at which *indeterminism* is true. Double's Negative Claim, $\sim\Diamond\exists x\exists y(Hx \mathbin{\&} Ay \mathbin{\&} Fxy \mathbin{\&} \sim Dxy$, is a view that we naturally might call "Secondary Free-will Incompossibilism", for it is the *denial* of Secondary Free-Will Compossibilism. Thus, Double's Negative Claim expresses an *incompossibility* thesis and is not an expression of (any type of) compossibilism or compatibilism. *Pace* Double, then, the claim that free will requires determinism is not a negative thesis of (primary) compatibilism, but is just a non-compatibilist thesis that a (primary) compatibilist may or may not endorse.

That said, I believe that James's description of soft determinism reveals that compatibilism does have a negative thesis—one that is rarely (if ever) discussed explicitly. The soft determinist, James tells us, is someone who *identifies* free will and determinism and who believes that "freedom is only necessity understood". I believe that James means for us to understand the soft determinist as someone who endorses (in addition to (P_A), (P_N), (P), and Secondary Incompossibilism) some type of necessary or "strict" compatibility principle. Paul Edwards, a self-identifying hard determinist (see below), attributes just such a principle to the soft determinist when he says that his soft determinist rivals hold that "there is in the first place *no contradiction whatsoever* between determinism and the proposition that human beings are sometimes free agents" (2002: 60; my emphasis). In other words, the soft determinism holds that there is *no possible world* at which determinism presents a threat to someone's freedom. I believe that we can express this strict non-contradiction principle as follows:

> Where we draw upon the above terms and where
> '$\Box$' represents (strong) metaphysical possibility

> *The Strict (Primary) Compatibility Thesis* (C): At no metaphysically possible world does there exist an agent who does not act freely *just because* her action is determined;
> $(C) =_{df} \Box\forall x\forall y((Hx \;\&\; Ay \;\&\; Dxy) \rightarrow {\sim}({\sim}Fxy\text{-}bc\text{-}Dxy))$.[11, 12]

Given that (C) is not a widely discussed compatibilist principle, it may seem strange at first glance. I will address this concern in Section 7. For now, I would like to note that the soft determinist is someone who endorses (C), each of three Primary Compossibility Theses, i.e. (P_A), (P_N), and (P), as well as Secondary Incompossibilism.

Let us now take stock of the views that we have discussed. As our discussion of the tenets of soft determinism has shown, one can consistently endorse (P_A), (P_N), (P), and (C). However, because a compatibilist need not hold that TD is true, the compatibilist need not endorse either (P_A) or (P_N). Furthermore, (C) and (P) are logically independent, so a philosopher in the Primary Free-Will (In)compatibility Debate may endorse one without endorsing the other. Since there is no entailment between (C) and any modal strain of Compossibilism, it will be useful to have a name for the theses which result from combining these theses:

> *The Actual Compossibility-Compatibility Thesis* (C_A), = the conjunction of (C) and (P_A)
> *The Nomological Compossibility-Compatibility Thesis* (C_N) = the conjunction of (C) and (P_N)
> *The Compossibility-Compatibility Thesis* (C_P) = the conjunction of (C) and (P)

In sum, the soft determinist endorses (C_A), from which it follows that the soft determinist endorses both (C_N) and (C_P)—which is to say that the soft determinist endorses every compatibility thesis we have discussed. However, as the soft determinist also endorses (at least

[11] Alternatively: ${\sim}\Diamond\exists x\exists y(Hx \;\&\; Ay \;\&\; Dxy \;\&\; ({\sim}Fxy\text{-}bc\text{-}Dxy))$.

[12] Technically, the conjunction of (C), TD, and DBT (each of which is endorsed by the soft determinist) also entails $\forall x\forall y((Hx \;\&\; Ay \;\&\; Dxy) \rightarrow {\sim}({\sim}Fxy\text{-}bc\text{-}Dxy))$ and $\boxed{N}\forall x\forall y((Hx \;\&\; Ay \;\&\; Dxy) \rightarrow {\sim}({\sim}Fxy\text{-}bc\text{-}Dxy))$. However, given that such qualified principles only make sense philosophically when (I) is presupposed, I believe that it would only cause unnecessary confusion to discuss these qualified theses in any detail. (The reader may consider the arguments I provide in Section 7 as a defense of my position on this matter.)

some expressions of) Secondary Incompossibilism, soft determinism cannot be *identified* with (C_A).[13]

The above discussion suggests that we have a choice to make about how we will use the term 'compatibilism'. Since compatibilist need not, *qua* being a compatibilist, believe either that determinism is true or that we have free will, 'compatibilism' cannot be defined plausibly in terms of either (P_A), (P_N), (C_A), or (C_N). However, it is not immediately obvious whether 'compatibilism' is best conceived of in terms of the compossibility thesis (P), the logical compatibility thesis (C), the conjunction of these theses (C_P), or, perhaps, as the disjunction of (C) and (P). Since the decision about which of these four options we should endorse partly depends on the logical relationships between (P), (C), and *incompatibilism*, let us put off our efforts to identify the one "real" compatibilism until after our discussion of its main rival.

4. Incompatibility and Incompatibilism

I believe that Vihvelin is right in thinking that we must not define 'incompatibilism' as the mere denial of (P) because we should not accept that Impossibilism entails Incompatibilism. Incompatibilism is not merely an answer to the question of whether the conjunction of TD and E-FWT_H might be true at some possible world. Incompatibilists do have a preferred answer to this

[13] In this essay, I describe soft and hard determinism as views only about the (in)compatibility of freedom and determinism. However, the reader might point out that James also indicates that the soft and hard determinist are interested in moral responsibility as well. For instance, speaking of soft determinism, James disparagingly says that this is "the determinism which allows considerations of good and bad to mingle with those of cause and effect" and the "dilemma of determinism" after which his talk is named has as a moral pessimism as one horn and subjectivism as the other (1956: 166). Of course, soft determinism *allows* (where hard determinism does not) that individuals may be morally responsible for their actions, but that does not make the thesis that we are morally responsible a *defining* tenet of soft determinism. In my opinion, James seems to discuss "optimism" and "pessimism" with respect to morality as *implications* of soft and hard determinism (respectively) rather than aspect of the views themselves. Perhaps I am wrong on this point. Either way, soft determinism is a complex view and van Inwagen was wrong to say that this view could be "easily defined" using merely his preferred terms, 'compatibilism' and 'determinism' (1986: 13).

question, of course: like the Impossibilist, the incompatibilist answers with a resounding "no".[14] However, incompatibilists also are committed to a positive thesis which distinguishes them from mere Impossibilists$_{\text{-H}}$—but *pace* Vihvelin, I deny that this additional thesis is Possibilism.

Incompatibilists and Impossibilists agree that (P) is false, but what sets them apart is that the incompatibilist (*qua* being an incompatibilist) forwards a particular view about *why* (P) is false whereas the mere impossibilist (*qua* being an impossibilist) does not. According to the incompatibilists, E-FWT$_{\text{H}}$ is false at every possible world at which TD is true *and* the truth of determinism at these worlds explains *why* E-FWT$_{\text{H}}$ is false at every world at which TD is true. So, while Incompatibilism is commonly misperceived as a strictly negative thesis (i.e. as the view that compatibilism, whatever that is, is false) we now see that the distinctive explanatory claim of Incompatibilism is its underappreciated *positive* thesis. In other words, among the defining tenets of Incompatibilism is:

> Where we draw upon the terms introduced above and
> where 'bc' represents "because"[15]
>
> *The Strict (Primary) Incompatibility Thesis* (I): Necessarily, anyone who is determined to perform an action is someone who does not freely perform that action *just because* her action is determined;
>
> (I) $=_{df}$ $\Box\forall x\forall y((Hx \;\&\; Ay \;\&\; Dxy) \rightarrow ({\sim}Fxy\text{-bc-}Dxy))$.[16,17]

[14] Of course, the impossibilist also asserts the stronger claim that E-FWT is false at every possible world at which TD is true.

[15] The compound sentence "A just because B" is true if and only if B provides a sufficient explanation for the truth of A.

[16]The fact that 'because' is not a truth-functional connective does not mean that there is something ill-formed about this statement of incompatibilism, nor does it imply that there is no fact of the matter whether this statement of incompatibilism is true or false. The operation posited in the formula "(~Fxy-bc-Dxy)" depends on something more the truth values of the terms 'Fxy' and 'Dxy' to determine the result of the operation, but there is a partial truth table for 'because' which is adequate for our purposes. As with a conjunction, if either '~Fxy' or 'Dxy' (or both) is false, then the compound sentence created with the connective 'because' is also false. However, the truth of the conjunction (~Fxy & Dxy) is not sufficient for the compound sentence ~Fxy-bc-Dxy to be true. In the case where the conjunction of these terms is true, additional work must be done to establish whether the 'because' claim is true or false. As such, a truth-value must be assigned to the whole sentence '~Fxy because Dxy' based on whether the sentence as a whole expresses something true or false. Furthermore, if the compound sentence "(~Fxy-bc-Dxy)" is true, it follows that "~Fxy" and "Dxy" are each true. So, just as the conjunction (*A* and *B*) entails *A*, what we

Clearly, (I) is the counterpart to the compatibilist thesis (C). By including a diagnostic “because” clause in (I) and identifying (I) as a defining tenet of incompatibilism, we ensure that no impossibilist who denies that determinism is a threat to free will shall qualify as an incompatibilist. So, Vihvelin was on the right path in thinking that Incompatibilism must be understood in terms of a positive thesis, she simply misidentified the positive thesis that Incompatibilists must endorse.

Incompatibilism is not plausibly defined merely as the view that (I) is true, however. I contend that there is one other modest thesis that incompatibilists must agree is at least metaphysically possibly true:

> Drawing on the terms introduce above and where
> ‘Hx’ represents *x is a human-like being* [18,19]
> ‘Dxy’ represents *x is determined by the laws to perform y* [20]
>
> *The Determined Human-like Being Thesis* (DBT): The conjunction of TD is true and some human-like being *x* performs an action *y*;
>
> (DBT)$=_{df}$ ∃x∃y(Hx & Ay & Dxy).[21]

might call the ‘bejunction’ (*A* because *B*) entails *A*. As such, the following formula is true based on the stables logic of ‘because’: **TRANS**: □∀x∀y ((~Fxy-bc-Dxy)→~Fxy). Using TRANS, one can give a formal proof (using the rules of S5) of every entailment claim made in this essay.

[17] Technically, the conjunction of (I), TD, and DBT also entails ∀x∀y((Hx & Ay & Dxy) → (~Fxy-bc-Dxy)) and $\boxed{N}$∀x∀y((Hx & Ay & Dxy) → (~Fxy-bc-Dxy)). However, as I explained above (see fn. 18), I believe that it would only cause unnecessary confusion to discuss these qualified theses in any detail. Again, the reader may consider the arguments I provide in Section 7 as a defense of my position on this matter.

[18] Clearly, ◊∃x(Hx) entails ◊∃x(Sx) and ◊E-FWT_H entails ◊E-FWT.

[19] In short, the “relevant” similarities are at least (1) the being cannot perform miracles with respect to the (natural or causal) laws or change the (natural or causal) laws and (2) the being possesses those general cognitive capacities by which we would group humans and fictional non-human entities like *Star Trek*’s Vulcans, Klingons, Romulans, Ferengi and (arguably) Data, into a common category of entities which seem to be candidates for free agency—e.g. capacity for second-order desires about first-order desires, the ability to weigh reasons for action, etc..

Notably, I purposefully avoid using the term ‘agent’ in the definition of ‘incompatibilism’. While I assume that it is uncontroversial that an being who performs a free action is also an agent, if ‘incompatibilism’ were defined in terms of agency, this would incorrectly express the logical commitments of the incompatibilist. In brief, an adequate definition of ‘incompatibilism’ will allow logical space for the incompatibilist who holds that determinism precludes *agency* and, so, *free agency* (see, for example, Helen Steward’s defense of “agency incompatibilism” in *A Metaphysics for Freedom,* forthcoming 2012).

[20] Notably, Dxy implies TD.

I use the phrase “human-like being” to refer to those (metaphysically possible) entities which are “human-like” in two critical ways. First, a human-like being is one that has roughly the same (or higher) degree of cognitive sophistication as a normal human being, meaning that it is an intelligent being which acts upon the basis of reasons, is capable of having second-order desires about its first-order desires, etc. Second, a human-like being is one that is subject to the natural laws of the universe in which it lives, meaning that such a being has no magical or “god-like” powers to change or to perform miracles with respect to the laws of nature (i.e. he does not have the power to “trump” the laws such that he can make things happen which would not have occurred as a function of the past together with the laws).[22] I would add that “human-like” is not meant to imply that the being must be a biological organism (an android, for instance, might be sufficiently human-like to be just as much a contender for free agency as we) nor does it imply that the being is a material substance (although, in this case, we must assume that there are some non-physical natural laws to which this immaterial being is subject).

At first glance, the fact that DBT discusses only “human-like beings” may seem at odds with my earlier critique of van Inwagen’s and Vihvelin’s preferred statements of the existential free-will thesis. However, when it comes to the Primary Free-Will (In)compatibility Debate, we are working within a very narrow context. In my view, Primary Free-Will (In)compatibility

[21] If TD is true, then any action performed by a human-like being is an action that he or she is determined by the laws to perform: $\square(TD \rightarrow \forall x \forall y(Hx \; \& \; Ay) \rightarrow (Dxy))$.

[22] For example, this second qualification seems to rule out the fictional *Star Trek: The Next Generation* character named “Q”. As described in the program, Q is cognitively sophisticated (thus satisfying the first condition of being ‘human-like’) and is (or is nearly) both omniscient and omnipotent. Q sometimes expresses his omnipotence (or so the story goes) by performing miracles and sometimes by changing the very laws of nature to suit his purposes. There are various theories one might suggest in an effort to explain how Q might be able to do this, but I do not mean to suggest that Q, thus described, is a metaphysically possible being. I wish only to point out that *if* Q is able to change the laws to which his states are subject, then he would clearly not be condemned to a particular future based on the facts of the past and the laws of nature which hold over any arbitrary period of time in his universe. As such, *if* such a being as Q is metaphysically possible, he is not the type of being which might be unfree *because* the laws of nature determine his future. Clearly, such a being as Q, then, is not the type of being under investigation in the Primary Free Will Debate.

Debate can be represented by two general questions: (1) “Does the predicate *free* apply to some action performed by a human-like being at some possible world at which determinism is true?” and (2) “If the answer to the first question is ‘no’, *why* is it ‘no’, and if the answer to the first question is ‘yes,’ why is it ‘yes’?”[23] There is logical space for a philosopher to claim that the answer to (1) is “no” because determinism is necessarily false. Clearly, though, this defense of a negative reply to (1) cannot be endorsed by someone who claims, as the incompatibilist does, that *the truth of determinism* at a world adequately explains *why* the predicate free has no application at that world. Likewise, if the incompatibilist were to deny that human-like beings exist at some possible world at which determinism is true, the incompatibilist would undercut his own proposal that the truth of determinism explains why no human-like beings are free at worlds where determinism is true—beings that do not exist do not perform actions and, so, *a fortiori* do not perform *free* action. Thus, I contend that incompatibilists *must* agree that DBT is true in at least one metaphysically possible world, or ‘$\Diamond$DBT’ for short.

Incompatibilism, then, is best characterized as the view that the conjunction of (I) and $\Diamond$DBT is true:

Incompatibilism $=_{\mathrm{df}}$ (I) & $\Diamond$DBT[24]

[23] Likewise, the Secondary Free Will Debate can be captured by two questions: (1) “Does the predicate *free* apply to some action performed by a human-like being at some possible world at which *indeterminism* is true?” and (2) “If the answer to the first question is ‘no’, *why* is it ‘no’, and if the answer is ‘yes’, why is it ‘yes’?”

[24] Notably, Strict-Incompatibility Incompatibilism entails Incompossibilism. That is, the conjunction of (I) and the (in)compatibility-neutral assumption $\Diamond$DBT entails (~P). Furthermore, the conjunction of (I), TD, and $\Diamond_N$DBT entails (~P_N), and the conjunction of (I), TD, and DBT entails (~P_A).[24] Now, since every incompatibilist must endorse (I) and $\Diamond$DBT, every incompatibilist endorses (P_IN) and (~P). All incompatibilists who endorse (I), TD, and $\Diamond_N$DBT, thereby also endorse (N_IN). Finally, all incompatibilists who endorse (I), TD and DBT thereby endorse (A_IN). Thus, while there is logical space for the compatibilist to endorse (C) and $\Diamond$DBT without endorsing compossibilism, there is no logical space for the incompatibilist to withhold endorsement of incompossibilism.

I believe that there may be some benefit in naming each of these bundles of incompatibilistic views, as they each represent a distinctly incompatibilist commitment to the incompossibility of free will and determinism. Naming each of the incompatibilistic bundles of views, we get:

Incompossibility-Incompatibilism (I_P) $=_{\mathrm{df}}$ (I) & (P_IN) & (~P) & $\Diamond$DBT.

Nomological Incompossibility-Incompatibilism (I_N) $=_{\mathrm{df}}$ (I) & (N_IN) & (~P_N) & $\Diamond_N$DBT.

In addition to being intuitively correct, the above definition leaves open an incompatibilist route to Impossibilism. In order to show that such a door must be left open, let us consider Max, a philosopher who endorses that necessarily, determinism precludes free will, i.e., (I). Max also holds that TD is necessarily true—perhaps because he is a law necessitarian who thinks that the actual laws are deterministic and/or because he believes that the notion of "indeterministic laws" is incoherent.[25] However, Max also denies the metaphysical possibility that some agent could perform a miracle with respect to the laws (or simply change the laws of nature) and, due to his cosmological and theological views, denies that there exists anything whatsoever beyond the boundaries of a given physical universe. Implied by this subset of Max's views is the thesis that there is no metaphysically possible world at which someone acts freely, or ~◊E-FWT. In short, Max is an Incompatibilist in virtue of endorsing (I) even though he is also an Impossibilist in virtue of endorsing ~◊E-FWT; Max is an Incompatibilist-Impossibilist.[26]

In addition, my preferred characterization of incompatibilism explains why libertarianism and hard determinism are each expressions of incompatibilism (and why, *pace* Doyle, it is not confusing to consider both libertarians and determinists "incompatibilists" (Doyle 2011: 61)). As James frames the free-will debate, it traditionally has been between what he calls "indeterminists" and "determinists", where the former championed our freedom and the latter argued against it. The indeterminists, James says, believe that they have the "sole right" to use

Actual Incompossibility-Incompatibilism (I_A) $=_{df}$ (I) & (A_{IN}) & ($\sim P_A$) & DBT.
Notably, there is some redundancy involved in expressing incompossibility-incompatibilism views in this way, for the first two conjuncts of each view entail the third conjunct. In this case, though, I do not find the redundancy problematic because it helps to illuminate that the existential incompatibility theses and incompossibility theses which constitute each modal strain of Incompossibility-Compatibilisms rise and fall together.

[25] Max may or may not believe that there are some possible worlds at which a universe exists which is non-law-governed. If *nothing* in such universes is law-governed, then it seems, at least *prima facie*, that no free agent exists in any such universe, for if there are *no* law-like connections—not even one-off laws—between the states of the agent, then it hardly seems that this being could have an enduring system for weighing reasons, etc. (For further discussion of the need for a certain type of law-like connections between the states of an agent, see my "Soft-Line Solution to Pereboom's Four-Case Argument" (2010) , which also appears as Chapter 2 of this dissertation).

[26] I borrow the term 'incompatibilist-impossibilist' from McKenna (2008: 443).

the term “freedom”, for they believe that freedom requires “variety” and “alternative possibilities” and such things cannot exist in a world at which determinism is true (James 1956: 149, 153). If we focus just on this commitment, James’s indeterminist is someone who is “soft” on indeterminism in just the way that the soft determinist is “soft” on determinism. The “Soft Indeterminist”, then, endorses a strict compatibility claim much like (C) but one whose focus is on indeterminism, namely: $\Box\forall x\forall y((Hx \ \& \ Ay \ \& \ {\sim}Dxy) \rightarrow {\sim}({\sim}Fxy\text{-bc-}{\sim}Dxy))$. In addition, the Soft Indeterminist also makes certain empirical claims. The Soft Indeterminist believes that some type of indeterminism is true (TD is false) and that we are (at least some of us, sometimes) free. So, we might express soft indeterminism as follows: There is no possible world at which determinism is true and some human-like being acts freely *because* necessarily, determinism precludes freedom; however, there are some possible worlds at which beings like us do act freely—namely, at some subset of the possible worlds at which indeterminism is true—and the actual world is one of these. Thus, the Soft Indeterminist endorses, *mutatis mutandis*, the same modal theses as the Soft Determinist.

Soft Indeterminism may seem to be equivalent to libertarianism. However, I think that libertarianism is better understood as subtype of Soft Indeterminism. Using my preferred language, libertarians typically argue from the assumption of Incompatibilism and the assumption that E-FWT$_H$ is true to the empirical claim that TD is false. In other words, libertarianism includes all the defining tenets of Soft Indeterminism, but libertarians also specify that a certain logical relationship holds between the defining tenets of Soft Indeterminism. The distinction between libertarianism and mere Soft Indeterminism, then, will be meaningful for those who accept that indeterminism is true and that we are free but do not think that it is reasonable to draw the conclusion that the laws of physics are indeterministic from the assumption that we are free.

Soft Indeterminists are, in their purely theoretical commitments, similar to James' "Hard Determinists". Like Soft Indeterminists, the Hard Determinist endorses (Primary) Incompatibilism. Unlike Soft Indeterminists, however, the Hard Determinist also holds that determinism (TD) is true and, so, concludes that E-FWT$_H$ is false. The Hard Determinist endorses:

> *The (Primary) Actual Incompatibility-Incompossibility Thesis* (A_{IN}): Determinism is true (at the actual world) and there exists some human-like being who is not free in performing some action *just because* he is determined to perform that action;
> $(A_{IN}) =_{df}\ \exists x \exists y\, (Hx\ \&\ Ay\ \&\ Dxy\ \&\ ({\sim}Fxy\text{-}bc\text{-}Dxy))$

In virtue of endorsing (A_{IN}), the Hard Determinist is also committed to:

> *The (Primary) Nomological Incompatibility-Incompossibility Thesis* (N_{IN}): Determinism is true (at the actual world) and there exists, at some nomologically possible world, a human-like being who is not free in performing some action *just because* he is determined to perform that action;
> $(N_{IN}) =_{df} \Diamond\!\!\!\!\!{\scriptscriptstyle N}\ \exists x \exists y\, (Hx\ \&\ Ay\ \&\ Dxy\ \&\ ({\sim}Fxy\text{-}bc\text{-}Dxy))$

In turn, (A_{IN}) and (N_{IN}) each entail:

> *The (Primary) Incompatibility-Incompossibility Thesis* (P_{IN}):
> At some metaphysically possible world, determinism is true and there exists a human-like being who is not free in performing some action *just because* he is determined to perform that action;
> $(P_{IN}) =_{df} \Diamond\ \exists x \exists y\, (Hx\ \&\ Ay\ \&\ Dxy\ \&\ ({\sim}Fxy\text{-}bc\text{-}Dxy))$

The mere Incompatibilist need not endorse (A_{IN}) or (N_{IN}). However, (P_{IN}) follows from the conjunction of ◊DBT and (I), so all Incompatibilists endorse (P_{IN}).

Notably, (P_{IN}), (N_{IN}), and (A_{IN}) are each existential theses, where each forwards a claim about *some* but not all beings whose actions are subject to deterministic natural laws. In other words, none of these three theses entails (I) (although the conjunction of ~(P_{IN}) and ◊DBT entails

~(I)). Furthermore, (P_{IN}), (N_{IN}), and (A_{IN}) are each logically consistent with (P), so one could endorse any of these thesis and still endorse Compossibilism. Thus, I contend that none of these three theses ((P_{IN}), (N_{IN}), nor (A_{IN})) is, properly speaking, an incompatibilist thesis. Still, insofar as each of these theses specifically appeals to the deterministic laws of nature to explain someone's lack of free will and each, if true, would entail that (C) is false, each is an "incompatibilistic" thesis.

The fact that (P_{IN}), (N_{IN}), and (A_{IN}) are logically consistent with (P), puts additional pressure on the view that compatibilism is adequately expressed by Compossibilism. Let us now turn to the issue of how Incompatibilism relates to Compossibilism.

5. (In)compossibilism

As discussed above, incompatibilism is commonly thought to be the mere denial of compatibilism, where the latter is understood as the view that (P) is true. However, as Vihvelin has shown us, the mere denial of (P) does adequately express incompatibilism. While Vihvelin, though, shows little interest in the view which results from the denial of (P), I think that we should keep track of any view which, if true, would entail that some popular conception of compatibilism is false. Such a view is surely interesting in its own right.

For each compossibility thesis (P), (P_N), and (P_A), there is a contradictory "incompossibility" thesis:

The Incompossibility Thesis (~P): $\sim\Diamond\exists x\exists y(Hx \mathbin{\&} Ay \mathbin{\&} Dxy \mathbin{\&} Fxy)$

The Nomological Incompossibility Thesis (~P_N): $\sim\Diamond_{N}\exists x\ \exists y\ (Hx \mathbin{\&} Ay \mathbin{\&} Dxy \mathbin{\&} Fxy)$

The Actual Incompossibilism Thesis (~P_A): $\sim\exists x\exists y(Hx \mathbin{\&} Ay \mathbin{\&} Dxy \mathbin{\&} Fxy)$

Looking at these theses, we see that they cannot be used to distinguish between the Impossibilists and Incompatibilists: Impossibilists will reject each thesis because they believe that the last conjunct of the thesis is necessarily false, while Incompatibilists will reject each thesis because they believe that the conjunction of the last two conjuncts is necessarily false. Where "Incompossibilism" is the view that The Incompossibility Thesis ~(P) is true, Vihvelin is surely right that Incompossibilism is not logically equivalent to Incompatibilism: Incompatibilism entails Incompossibilism but Incompossibilism does not entail Incompatibilism.

6. Compatibilism

Compatibilism is typically thought of as a positive thesis and incompatibilism as the denial of that positive thesis, whatever it is, and vice versa. Looking at Compossibilism, we see the positive view that is typically identified with compatibilism, and with Incompossibilism we see the negative view that is typically identified with incompatibilism. Since Compossibilism entails and is entailed by the negation of Incompossibilism, we can understand why there is a common presumption that compatibilism and incompatibilism are contradictories. We have seen, though, that it is a mistake to think of compatibilism as a merely positive thesis; compatibilism is not logically equivalent to Compossibilism. We have seen also that incompatibilism is not a merely negative thesis; incompatibilism is not logically equivalent to Incompossibilism. Moreover, we have seen that one need not endorse Compossibilism in order to reject the unique positive thesis of Incompatibilism. This means that one can reject Incompatibilism regardless of whether one rejects Incompossibilism. This also means that there is logical space for one to take up an "anti-incompatiiblist" stance by endorsing the negative Strict Compatibility Thesis (C) without endorsing the positive Compossibility Thesis (P). What, then, is to be made of the modest anti-incompatibilist view that (C) is true?

As suggested by my terminology, I believe that the Strict Compatibility Thesis (C) is the defining tenet of compatibilism. In saying this, I do not mean that (C) is *among* the defining tenets, but that it is the *only* defining tenet of compatibilism. I expect that most readers will balk at this minimalist conception of compatibilism. Most likely, those who disapprove will do so because most practicing compatibilists endorse (P) and because the view that compatibilism is adequately expressed by (P) is so deeply entrenched. Admittedly, (P) has the right pedigree to be a defining tenet of compatibilism—the free-will debate arose from the compossibilist assertion by the Stoics that determinism is true and that we are free (cf. Bobzien 1998). These considerations make it seem that my suggested definition of 'compatibilism' is overly revisionary. However, since the standard definition of 'compatibilism' fails to capture the view intuitively associated with the term, some revision to the standard is required. that we do not think that 'compatibilism' is logically equivalent to Soft Determinism just because the first compatibilists were also Soft Determinists. Likewise, the descriptive fact that most practicing compatibilists endorse (P) does not, on its own, imply that compatibilism is logically equivalent to Compossibilism.

Indeed, the claim that compatibilism is logically equivalent to Compossibilism is demonstrably false. As noted above, (P) and (P_{IN}) are logically consistent theses.[27] Thus, the truth of (P) alone would not rule out the possible existence of some human-like being who is not free just because his actions are determined by the natural laws—but, intuitively, compatibilism does! Presumably, the most substantial piece of evidence in favor of the view that (P) is the only defining tenet of compatibilism is the descriptive fact that most self-identifying compatibilists endorse (P). With that in mind, I raise the question: Who among the self-identifying compatibilists would, having shown (I) to be false, be inclined to allow the crippled incompatibilists to take permanent refuge in the incompatibilistic theses (P_{IN}), (N_{IN}), or (A_{IN})?

[27] If the 'x' in each refers to the same object then a contradiction does arise. However, the 'x' in (P) and the 'x' in (P_{IN}) are not under the scope of the same existential quantifiers, so 'x' may represent a different object in each thesis.

Not only do I doubt that any self-identifying compatibilist would be inclined to offer quarter to P_{IN}-, N_{IN}-, or A_{IN}-theorists, but I also contend that it would be philosophically suspect for such quarter to be given. (P) and (P_{IN}) are logically consistent, but consider what happens when the Compossibilist and the mere P_{IN}-theorist are asked to comment upon the freedom of a particular individual living in a deterministic universe. Let us say that Cain and Abel live in a deterministic universe and Cain kills Abel and that 'x' represents Cain and 'y' represent Cain's act of killing Abel both in (P) and in (P_{IN}). According to (P), the proposition "Cain freely kills Abel" is true, but according (P_{IN}) this proposition is false. Assuming that the P_{IN}-theorist and the Compossibilist are neither talking nonsense nor talking past each other, one of these philosophers is rightly describing Cain and the other is not.[28] Where does the Compossibilist go from here?

I contend that the Compossibilist must squarely face off with his opponent's *positive explanation* for why Cain is not free and *categorically* reject it. If the Compossibilist does not do this, she will face the same formal battle anew with regard to every individual action of any given human-like being at any possible world at which TD is true. In order to quiet her adversary permanently, the Compossibilist must demonstrate more than the truth of (P); she must argue that one's being subject to deterministic laws is *never* sufficient to undermine a person's free will. Those who accept this task take up the burden of the compatibilist—and this is the task of defending (C).

Of course, denying that (P) is a defining tenet of compatibilism does not prevent us from acknowledging that most philosophers who are self-identifying compatibilists are deeply concerned with the truth of (P). Surely, most practicing compatibilists are committed to at least

[28] In other words, we are assuming that free-will non-cognitivism is false and that the two philosophers are using the term 'freely' to express the same concept. Notably, even if one were to endorse some type of radical free-will relativism where one is only as free as she feels, there would be a fact of the matter with respect to which of the two philosophers correctly describes the freedom-status of Abel.

(C_P), the view that I have called Compossibility-Compatibilism, which is the view that *both* (P) and (C) are true. However, most practicing compatibilists probably also endorse (C_A), which is to say that most practicing compatibilists believe that we have free will even if—or perhaps *because*—the laws of our universe are deterministic. Still, I contend that the debate between compatibilists and incompatibilists boils down to a disagreement over a necessity claim rather than a possibility claim about whether (C) or (I) is true.

7. The Logical Gap Observed

Recognizing the various expressions of Incompossibilism might seem interesting only insofar as these views help us to complete our picture of the logical space in which the Primary Free-Will Debate takes place. However, I believe that Incompossibilism plays a larger role in the free literature than one might think. Consider, for instance, an argument strategy that McKenna has dubbed “The Manipulation Argument”. According to McKenna, this argument is an argument for the incompatibility of determinism and free will (and, so, moral responsibility), describing the argument as follows:

> [The Manipulation Argument] involves manipulation of an agent. It is, really, an argument form, and different instances of it are formulated around different examples and different compatibilist accounts of free will. Roughly, the argument begins with an example of an agent manipulated in manner M into (allegedly) satisfying compatibilist sufficient conditions for free will (and moral responsibility), CSC. The agent then performs an act as a causal upshot of CSC. The case is supposed to elicit the thought that, owing to the manipulation, the agent does not act freely (and is not morally responsible).
> Here is the basic form of The Manipulation Argument (MA):
> 1. Any agent manipulated in manner M into satisfying CSC does not act freely (exercise her free will).
> 2. Determinism is in no relevant manner any different from M—it is just a different way to bring about CSC.
> 3. Therefore, acting freely is incompatible with determinism; CSC is insufficient for free will. (McKenna 2008: 439)

While McKenna's claim that MA is a template for an argument for incompatibilism, it is unclear whether this template provides anything beyond an argument for Incompossibilism. The confusion arises with Presmise 2. Should Premise 2 be understood as the positive diagnostic claim that the manipulation is freedom-undermining because it involves deterministic causation or, instead, as the negative claim that there is no freedom-relevant difference between someone who is manipulated to perform an action and being determined to perform that action? If the latter, the MA is not a template for Incompatibilism—Impossibilists who reject (I) can nonetheless agree that there is no freedom-relevant difference between these scenarios. If the former, MA is a template for Incompatibilism, but the important argument is hidden from view: the real work of MA is now being done by the sub-argument which *supports* the positive diagnosis asserted in Premise 2.

In order to see the problem with MA in greater relief, let us consider a paradigm instance of this template: Pereboom's famous "Four-Case Argument".[29] As Pereboom describes his argument, it proceeds in two stages, "a combined counterexample and generalization strategy", to the ultimate conclusion that there is no set of compatibilist-friendly sufficiency conditions for moral responsibility. Pereboom claims that his argument has an "incompatibilistic" conclusion (2001: 112). Let us take Pereboom's own description of his argument seriously. Seen as the mere combination of a counterexample and a generalization strategy, the success of the Four-Case Argument does not depend on a correct diagnosis of what specific feature of the manipulation described in case one does the work of undercutting the victim's freedom. The success of the counterexample depends on there being *some* feature of the manipulation story which does, in fact, undercut the freedom and responsibility of the victim, while the success of the

[29] Indeed, McKenna's formulation of The Manipulation Argument was originally introduced as a formal representation of the Four-Case Argument (McKenna 2004).

generalization strategy requires only that this feature—whatever it is—generalizes to the normal (determined) agent described in the fourth and final case. If both the counterexample and the generalization strategy work, then Pereboom succeeds in showing that at no metaphysically possible world at which the causal laws are deterministic and there exists some human-like being who satisfies the sufficiency conditions for freedom and moral responsibility. In other words, as Pereboom describes his argument, the Four-Case Argument is attack on $\Diamond\exists x\exists y(Hx \ \& \ Ay \ \& \ Dxy \ \& \ Fxy)$, a.k.a., Compossibilism. However, as we have seen, the mere denial of compossibilism is not incompatibilism, but *Incompossibilism*.

Amid his presentation of the four cases, Pereboom also suggests that the *best explanation* for why each of the (purported) victims in his four cases is free and responsible: in each case, the victim’s actions are causally determined by factors beyond his control. Later, in a section after his presentation of the Four-Case Argument, Pereboom claims that the argument gives us “good reason to believe that an agent cannot be responsible for decisions that are produced by a deterministic process that traces back to causal factors beyond her control” (2001: 126). Now, were it true that determinism is the freedom-undermining feature in each case, then it would be true that each of the victims in Pereboom’s four cases is not free and this is *because* each is determined to do as they do. So, seen as a best-explanation argument, the Four-Case Argument does seem to be an argument for Incompatibilism.

While some philosophers believe that the Four-Case Argument can be reduced to a best-explanation argument for Incompatibilism (cf. Mele 2008), the standard view is that the essential structure of the Four-Case Argument is that described by MA where Premise 2 is understood as a mere no-difference claim. I contend that there are at least four good reasons to favor the latter over the former view of Pereboom’s argument. First, Pereboom introduces the Four-Case

Argument as a two-stage argument against the view that there is a set of compatibilist-friendly sufficient conditions for free and responsible agency. As noted above, such a conclusion entails incompossibilism but not incompatibilism. Second, there is logical space for both stages of the Four-Case Argument (against Compossibilism) to succeed even if determinism is *not* the correct explanation for the victims' lack of freedom and moral responsibility.[30] Finally, Pereboom makes such a minimal effort to identify and rule out plausible alternatives to his preferred diagnosis of the freedom-undermining feature of the manipulation that one is hard-pressed to see an *argument* to the best explanation within the Four-Case Argument. In the light of these points, Pereboom's appeal to determinism is most naturally seen as mere support for a premise in his generalization argument (insofar as the best way for Pereboom to support the truth his claim that there is no morally relevant difference between his four cases is for him to identify the responsibility-undermining feature that that the cases have in common). More specifically, Pereboom's best-explanation proposal is best seen as support for an instance of Premise 2 of the Manipulation Argument.

Viewed as an instance of the Manipulation Argument, it is less perplexing that Pereboom does not develop a robust argument to the best explanation in defense of his proposed diagnosis of the freedom- and responsibility-undermining feature of the manipulation cases. Seen as auxiliary support for a premise of the Four-Case Argument, Pereboom's best-explanation proposal lends strength to his argument at no risk. Pereboom need not make a serious effort to rule out all alternative explanations because the conclusion of his argument—seen as an instance of MA—does not depend on his diagnosis being correct.

[30] This means that even if Mele's critique (Mele 2008) of Pereboom's argument to the best explanation succeeds, Mele does not thereby show that the Four-Case Argument is unsound. Given that most compatibilists endorse compossibilism, the Four-Case Argument would continue to be one of the most important arguments in contemporary free will literature even if it were universally agreed that it is an argument for incompossibilism rather than incompatibilism.

Of course, one might wonder: If not determinism, then what *is* the common feature? Well, for all Pereboom says, perhaps the unhappy lesson of the Four-Case Argument is that one cannot be *created* by forces beyond her control and still be free—*regardless* of whether she is created in a deterministic or indeterministic world. Beyond Pereboom's *suggestion* that determinism is the freedom-undermining feature of the manipulation, though, there is nothing in the Four-Case Argument which indicates that *determinism* is the reason that none of his victims are free or responsible. In the light of these considerations, I contend that the Four-Case Argument was not designed to be an argument for any strain of Incompatibilism.

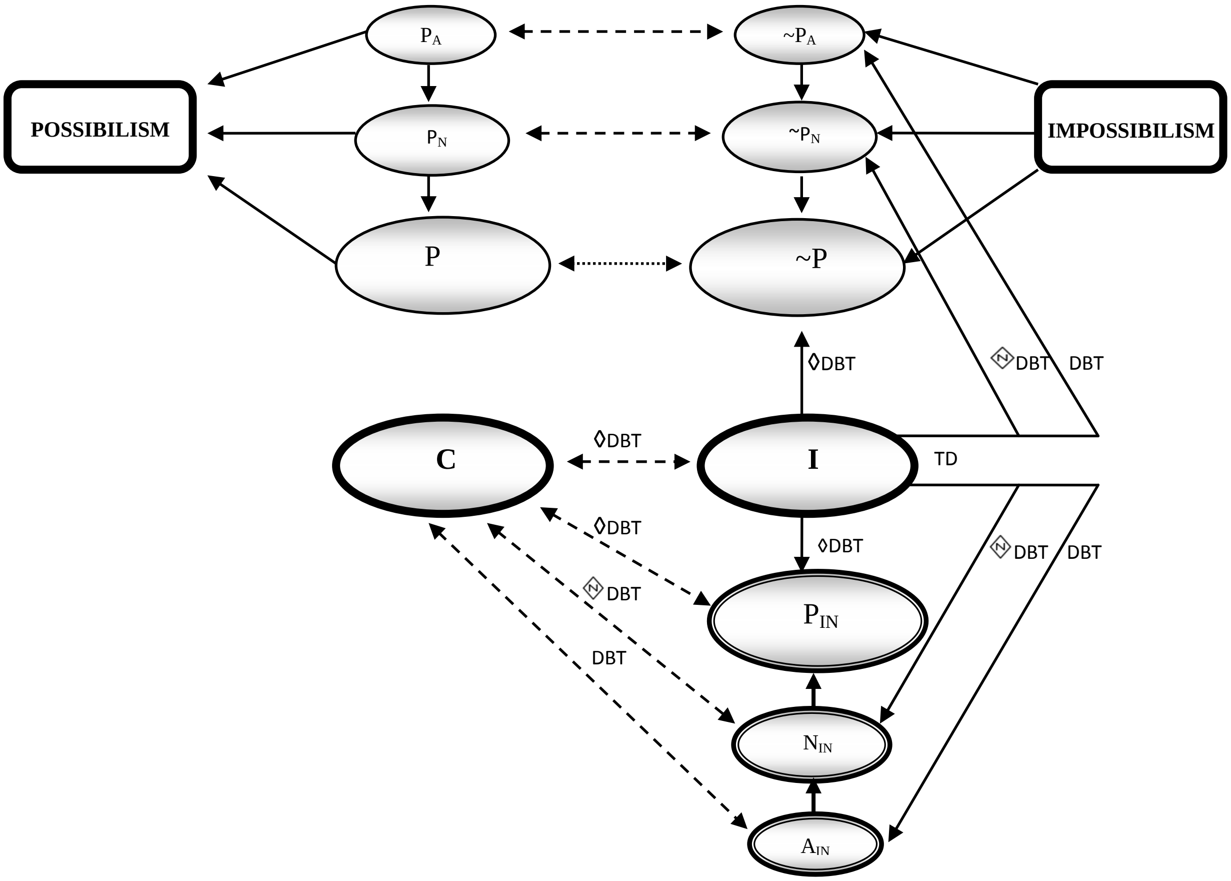
Of course, securing the conclusion that the Four-Case Argument fails as an argument for Incompatibilism on the grounds that it succeeds as an argument against Compossibilism would be a pyrrhic victory for most Compatibilists, since most endorse Compossibilism. However, whether some group of philosophers is *personally satisfied* by a certain reply to the Four-Case Argument or whether the Four-Case Argument is a threat to some popular and dearly-held view is beside the point, since I am not suggesting that we must accept that the Four-Case Argument is a *sound* argument for Incompossibilism—I have argued elsewhere that it is not (Demetriou 2010). My goal in discussing the Four-Case Argument is only to highlight how easily an argument for mere Incompossibilism is mistaken for or conflated with an argument for Incompatibilism. Moreover, by showing that the paradigm instance of MA does not rise to its billing as an argument for Incompatibilism, our review of the structure of the Four-Case Argument also reveals that the logical structure of the MA is ill-defined and does not clearly represent a class of arguments for Incompatibilism. Contrary to popular belief, that is, MA is designed to be a template for a class of arguments for Incompossibilism rather than Incompatibilism.

8. Closing Remarks

In this essay, I have surveyed the logical space in which the Free-Will (In)compatibility Debates take place. In doing this, I have revealed that even the most familiar views in the free will debate are poorly understood and inadequately articulated. The "Modal Map" that I provide below offers a summary of the major theses and views discussed in this essay and the important logical relationships between them.

THE MODAL MAP:

COMPATIBILISM	**INCOMPATIBILISM**
(C): $\Box\forall x\forall y((Hx \,\&\, Ay \,\&\, Dxy)\rightarrow \sim(\sim Fxy\text{-bc-}Dxy))$	**(I) =** $[\Box\, \forall x\forall y((Hx \,\&\, Ay \,\&\, Dxy) \rightarrow (\sim Fxy\text{-bc-}Dxy))] \,\&\, \Diamond DBT$
COMPOSSIBILISMs	**INCOMPOSSIBILISMs**
(P) $=_{df}\ \Diamond\ \exists x\exists y\ (Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$ **(P_N)** $=_{df}\ \Diamond_N \exists x\exists y\ (Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$ **(P_A)** $=_{df}\ \exists x\exists y\ (Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$	**(~P) =** $\sim\Diamond\exists x\exists y(Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$ ($\sim P_N$) = $\sim\Diamond_N\exists x\ \exists y\ (Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$ ($\sim P_A$) = $\sim\exists x\exists y(Hx \,\&\, Ay \,\&\, Dxy \,\&\, Fxy)$
COMPOSSIBILITY-COMPATIBLISMs	**INCOMPATIBILIST-INCOMPOSSIBILISMs**
(C_P) = (C) & (P) **(C_N)** = (C) & (P_N) **(C_A)** = (C) & (P_A)	**(I_P)** = (I) & (~P) & (P_{IN}) -Where **(P_{IN})** = $\Diamond\exists x\exists y\ (Hx \,\&\, Ay \,\&\, Dxy \,\&\, (\sim Fxy\text{-bc-}Dxy)$ **(I_N)** $=_{df}$ (I) & (~N) & (N_{IN}) -Where **(N_{IN})** = $\Diamond_N\exists x\exists y\ (Hx \,\&\, Ay \,\&\, Dxy \,\&\, (\sim Fxy\text{-bc-}Dxy)$ **(I_A)** $=_{df}$ (I) & (~A) & (A_{IN}) -Where **(A_{IN})** = $\exists x\exists y\ (Hx \,\&\, Ay \,\&\, Dxy \,\&\, (\sim Fxy\text{-bc-}Dxy)$
DBT = $\exists x\exists y(Hx \,\&\, Ay \,\&\, Dxy)$ **TD =** The Thesis of Natural Law Determinism; $(TD\rightarrow(\forall x\forall y((Hx \,\&\, Ay) \rightarrow Dxy))$; $\Box\forall x\forall y((Dxy)\rightarrow TD)$	
‘$\Diamond_N$’ represents nomological possibility; ‘◊’ represents strong metaphysical possibility ‘ ⟶ ’ represents logical entailment; ‘ ◄- - -► ’ connects contrary views; ‘ ◄·······► ’ connects contradictory views	



BIBLIOGRAPHY

Beebee, Helen, and Alfred Mele 2002. Humean Compatibilism, *Mind* 111: 201-24.

Bird, Alexander 2004. Strong Necessitarianism: The Nomological Identity of Possible Worlds, *Ratio* XVII: 256-76.

Bobzien, Susanne 1998. *Determinism and Freedom in Stoic Philosophy.* New York: Oxford University Press..

Broad, C.D. 1934. *Determinism, Indeterminism, and Libertarianism*. Cambridge: The University Press.

Byrne, Patrick H. 1981. Relativity and Indeterminism, in *Foundations of Physics* 11.12: 913-32.

Campbell, Joseph Keim 2007. Free Will and the Necessity of the Past, *Analysis* 67.2: 105–11.

Craig, Edward ed. *Routledge Encyclopedia of Philosophy*, Vol 10.

Demetriou, Kristin 2010. The Soft-Line Solution to Pereboom's Four-Case Argument", *Australasian Journal of Philosophy* 88.4: 595-617.

Double, Richard 2002. *"Metaethics, Metaphilosophy, and Free Will Subjectivism", in Free Will, ed. Robert Kane. Massachusetts: Blackwell Publishers: 506-528.*

-----1996. *Metaphilosophy and Free Will*. New York: Oxford University Press.

Doyle, Bob 2011. *Free Will*, Massachusetts: I-Phi Press.

Edwards, Paul 2002. Hard and Soft Determinism, in *Free Will*, ed. Robert Kane. Massachusetts: Blackwell Publishers: 59-67.

Feeney, Stephen *et al* 2010. First Observational Tests of Eternal Inflation, http://arxiv.org/abs/1012.1995.

Fischer, John Martin 2004. Responsibility and Manipulation, *The Journal of Ethics* 8/2: 145–77.

-----1994. *The Metaphysics of Free Will*: *An Essay on Control*. Oxford: Blackwell.

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

Frankfurt, Harry 1969. Alternate Possibilities and Moral Responsibility, *The Journal of Philosophy* 66.23: 829–39.

Geach, P.T. 1961. Aquinas, in *Three Philosopher*s, eds. G.E.M. Anscombe and P.T. Geach. New York: Cornell University Press: 65-125.

Ginet, Carl 1990. *On Action*. Cambridge: Cambridge University Press.

Haji, Ishtiyaque and Stefaan Cuypers 2006. Hard- and Soft-Line Responses to Pereboom's Four-Case Manipulation Argument, *Acta Analytica* 21/4: 19–35.

Hawking, Stephen and Leonard Mlodinow 2010. *The Grand Design*. New York: Bantam Books.

Hobart, R.E. 1934. Free Will as Involving Determination and Inconceivable without It, *Mind*, 63: 1–27.

Hoefer, Carl 2010. Causal Determinism, in *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*, Edward N. Zalta (ed.),
URL = http://plato.stanford.edu/archives/spr2010/entries/determinism-causal/.

Honderich, Ted 2002. Determinism and the Real Problem, in *The Oxford Handbook of Free Will*, ed. Robert Kane, New York: Oxford University Press: 461-76.

Huemer, Michael 1996. The Subjectivist's Dilemma, *Objectivity* XX: 77-92.

James, William, 1956. The Dilemma of Determinism, *in The Will to Believe and Other Essays in Popular Philosophy*. New York: Dover Publications: 145–183. (Originally presented as a public address; first published as "An Address to the Harvard Divinity Students" in the *Unitarian Review*, September 1884.*)*

Kane, Robert 2002. Glossary, in *Free Will*, ed. Robert Kane. Massachusetts: Blackwell Publishers: 284-290.

Kim, Jaegwon 2005. *Physicalism, Or Something Near Enough.* Princeton: Princeton University Press.

-----1998. *Mind in a Physical World*. Cambridge, MA: The MIT Press.

-----1993. The Nonreductivist's Troubles with Mental Causation, in *Supervenience and Mind*. Cambridge: Cambridge University Press: 336–57.

Kneale, William, "Natural Laws and Contrary-to-Fact Conditionals", in *Analysis*, vol. 10, no. 6 (June 1950), pp. 121-125.

Laplace, Pierre Simon 1820/1951. *A Philosophical Essay on Probabilities*, New York: Dover Publications (translated from the 6th French edition by Frederick Wilson Truscott and Frederick Lincoln Emory).

Leslie, John 1989. *Universes,* New York: Routledge.

Linde, Andrei 1994. The Self-Reproducing Inflationary Universe, *Scientific American*, Vol. 271, No. 5: 48-55.

McGinn, Colin 1993. *Problems in Philosophy: The Limits of Inquiry*. New York: Wiley Blackwell.

McKenna, Michael 2010. Whose Argumentative Burden, which Incompatibilist Arguments?—Getting the Dialectic Right, *Australasian Journal of Philosophy,* 88.3: 429-443.

-----2008. A Hard-Line Reply to Pereboom's Four-Case Argument, *Philosophy and Phenomenological Research* 77.1: 142–59.

-----2004. Responsibility and Globally Manipulated Agents, *Philosophical Topics* 32: 169–92.

Mele, Alfred 2009. Free Will, in *Encyclopedia of Consciousness, Vol. 1*, ed. William P. Banks, Academic Press, Elsevier Inc.: 265-278.

-----2008. "Manipulation, Compatibilism, and Moral Responsibility", *Journal of Ethics* 12: 263-86.

-----2006. *Free Will and Luck,* New York: Oxford University Press.

-----1995. *Autonomous Agents,* New York: Oxford University Press.

Nahmias, Eddy and Dylan Murray Nahmias 2011 (forthcoming). Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions, in *New Waves in Philosophy of Action*, eds. J. Aguilar, A. Buckareff, and K. Frankish. England: Palgrave-Macmillan.

Nahmias, Eddy, Justin Coates and Trevor Kvaran 2007. Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions, *Midwest Studies in Philosophy* XXXI: 214-242.

Nozick, Robert 1981. *Philosophical Explanations*. Cambridge: Harvard University Press.

Olson, Jonas 2010. In Defence of Moral Error Theory, in *New Waves in Metaethics*, ed. Michael S. Brady. Palgrave Macmillan. Palgrave Connect. Palgrave Macmillan. 18 Nov 2011 http://www.palgraveconnect.com/pc/doifinder/10.1057/9780230294899

Penrose, Roger *et al* 2010. Concentric Circles in WMAP Data May Provide Evidence of Violent Pre-Big-Bang Activity, http://arxiv.org/abs/1011.3706.

Pereboom, Derk 2009. Hard Incompatibilism and Its Rivals, *Philosophical Studies* 144: 21–33.

-----2002. Living Without Free Will: The Case For Hard Incompatibilism, in *The Oxford Handbook of Free Will*, ed. Robert Kane, New York: Oxford University Press: 477–88.

-----2001. *Living Without Free Will*, Cambridge: Cambridge University Press.

-----1995. Determinism al Dente, *Noûs*, 29.1: 21–45.

Rees, Martin 1997. *Before the Beginning: Our Universe and Others*. Massachusetts: Perseus Books.

Roskies, Adina 2006. Neuroscientific Challenges to Free Will and Responsibility, *Trends in Cognitive Sciences* 10.9: 419-423.

Rummens, Stefan and Stefaan Cuypers 2010. Determinism and the Paradox of Predictability, *Erkenntnis* 72: 233–249.

Schrenk, Markus 2005. The Bookkeeper and the Lumberjack. Metaphysical vs. Nomological Necessity, in G. Abel (ed.), *Kreativität. XX. Deutscher Kongress für Philosophie. Sektionsbeiträge Band 1.* Universitätsverlag der Technischen Universität: ??.

Schnieder, Benjamin 2011. A Logic for 'Because', *The Review of Symbolic Logic* 4.3: 445-465.

Sehon, Scott 2010. A Flawed Conception of Determinism in the Consequence Argument, *Analysis* 71.1: 30-38.

Shabo, Seth 2011. What must a proof of incompatibilism prove?, *Philosophical Studies* 154: 361–71.

Smilansky, Saul 2007. Determinism and Prepunishment: the radical nature of compatibilism, *Analysis* 67.4: 347-349.

Sommers, Tamler 2006. *Beyond Freedom and Resentment: An Error Theory of Free Will and Moral Responsibility*. Dissertation <<ProQuest Information and Learning Company, UMI Microform 3228152>>

Strawson, Galen 1986. *Freedom and Belief*. Oxford: Clarendon Press.

–– 1994. The Impossibility of Moral Responsibility, *Philosophical Studies*, 75: 5–24. Reprinted in Watson 2003.

–– 2002. The Bounds of Freedom, in *The Oxford Handbook of Free Will, ed. Robert Kane*. Oxford: Oxford University Press: 441–60.

Steward, Helen 2012 (forthcoming). *A Metaphysics for Freedom*. New York: Oxford University Press.

van Fraassen, Bas 1995. `World' Is Not a Count Noun, *Noûs* 29.2: 139-157.

van Inwagen, Peter 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

-----1980. Indexicality and Actuality, *The Philosophical Review* 89.3: 403-426.

Vihvelin, Kadri 2011. Arguments for Incompatibilism. In *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition (first archived Winter 2003 Edition))*, ed. Edward N. Zalta, URL = <http://plato.stanford.edu/archives/spr2011/entries/incompatibilism-arguments/>.

—— 2008. Compatibilism, Incompatibilism, and Impossibilism. In *Contemporary Debates in Metaphysics*, eds. John Hawthorne, Theodore Sider, and Dean Zimmerman, 303–18. Malden, MA: Blackwell Publishing.

Warfield, Ted A. 2000. Causal Determinism and Human Freedom are Incompatible: A New Argument for Incompatibilism, *Philosophical Perspectives* 14, *Action and Freedom*: 167–80.

Zimmerman, David 1999. Born Yesterday: Personal Autonomy for Agents without a Past, *Midwest Studies in Philosophy* 23: 236–66.